Italian Journal of Animal Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjas20

Identifying the optimal number of topics in text mining: a case study on reindeer pastoralism literature

Barbara Contiero, Øystein Holand & Giulio Cozzi

To cite this article: Barbara Contiero, Øystein Holand & Giulio Cozzi (2024) Identifying the optimal number of topics in text mining: a case study on reindeer pastoralism literature, Italian Journal of Animal Science, 23:1, 1348-1357, DOI: 10.1080/1828051X.2024.2398168

To link to this article: https://doi.org/10.1080/1828051X.2024.2398168

Published online: 04 Sep 2024.

ITALIAN JOURNAL OF ANIMAL SCIENCE
2024, VOL. 23, NO. 1, 1348–1357
https://doi.org/10.1080/1828051X.2024.2398168

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Identifying the optimal number of topics in text mining: a case study on reindeer pastoralism literature

Barbara Contiero[a], Øystein Holand[b] and Giulio Cozzi[a]

[a]Department of Animal Medicine, Production and Health, University of Padua, Legnaro, Italy; [b]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, NMBU, Ås, Norway

ABSTRACT

Text mining and topic analysis algorithms which group textual contents in the most efficient way, are becoming increasingly useful to summarise the main information contained in large data corpus of complex scientific fields. Using the literature about reindeer pastoralism as a case study, this methodological investigation addressed the issue related to the identification of the suitable number of topics that provide the best in-depth interpretation of a large data corpus. Two-thousand eight hundred and seventy-five documents extracted from Scopus® regarding the scientific literature of reindeer pastoralism were used. Four simulations with 8, 10, 12, and 20 topics were carried out to define the optimal number of topics that best explained the issues related to reindeer husbandry. The results showed that a reasonable trade-off between the number of articles and the number of topics, based on the reduction of the variance explained within the group, leads to an optimal choice in the search for the most meaningful simulation. The adoption of a too large number of topics, with the excessive fragmentation of the data corpus into small aggregations of documents, encourages the emergence of topics without any technical or practical meaning, solely as a result of the unsupervised iterative process.

HIGHLIGHTS

- Text mining for insight vast and complex scientific fields: a case study on reindeer pastoralism.
- Optimising topic identification to strike a balance between the size of the articles corpus and the number of topics and achieve the most insightful results.
- Too many topics can lead to fragmentation and irrelevant results, while too few may over-simplify the complexity of the dataset.

ARTICLE HISTORY
Received 5 May 2024
Revised 9 August 2024
Accepted 24 August 2024

KEYWORDS
Number of topics; reindeer pastoralism; simulations; text mining; topic analysis

Introduction

The scientific literature has surged exponentially with a growth rate of 8–9% per year over the past decades (Bornmann and Mutz 2015). With the number of new scientific papers exceeding 5.0 million in 2022, the scientific communities are easily overloaded with information. As a response, several methods have been developed to navigate and handle this massive flood of publications. Systematic literature reviews are useful tools for understanding the state of the art of a given scientific topic as well as for propelling further research on it (O'Connor and Sargeant 2015). However, the target to extract and understand the main information is becoming increasingly complex and time-consuming when dealing with scientific fields that involve large collection of documents. The text mining and topic modelling analyses represent suitable alternatives for lightening the burden associated with document screening as they produce a fully unsupervised, structured 'map' of textual knowledge (Wang et al. 2016; Park and Kremer 2017). Text mining and topic analysis can be used to: summarise or cluster information from a large data corpus into charts or maps; identify hidden associations between concepts or textual elements; provide an overview of the contents of a large collection of documents; categorise articles or texts by discovering relevant groupings. For these reasons, they have been increasingly used in the scientific literature analyses as tools to identify themes and future research avenues (Antons et al. 2020; Nalon et al. 2020; Zuliani et al. 2021). Clustering

CONTACT Barbara Contiero barbara.contiero@unipd.it

procedures rely on unsupervised algorithms that group textual content based on similarity among texts (Jain 2010) and are part of the broader category of dimensionality-reduction techniques. This approach takes words into account and clusters them based on the assumption that terms that are co-present in the documents are used to express a certain latent topic. Based on those topics, document similarity can be quantified as the probability of membership in a topic (Antons et al. 2020).

This article addresses a methodological issue related to the use of text mining and topic analysis, namely the identification of the suitable number of topics that best summarise the information of a large data corpus, using the literature about reindeer pastoralism as a case study. Indeed, reindeer pastoralism is a complex socio-ecological production system that is practiced throughout the northern taiga and tundra of the Eurasian continent by an array of ethnic groups. A wide, complex, and multidisciplinary scientific literature has addressed reindeer pastoralism, embracing environmental, production, socio-cultural, and historical issues, which have shown a growing trend in terms of transdisciplinary studies over the last decades (Pape and Löffler 2012). The main findings of the analysis have been previously documented (Holand et al. 2024). However, in the present article, we aim to elucidate and clarify the challenges faced in terms of the practical implementation of the method, which is already extensively utilised across various fields (Wang et al. 2016: Brscic et al. 2021), aquaculture (Tucciarone et al. 2024), livestock and companion animals (Adamaκopoulou et al. 2023; Benedetti et al. 2023; Masebo et al. 2023; Trapanese et al. 2024), specifically focusing on the assessment of the optimal number of groups (topics) required to best characterise a collection of scientific articles. The rationale behind this approach is that if a too small number of topics is assumed, the complexity of a corpus may not be well interpreted. On the other hand, the choice of too large number of topics can lead to difficulties in interpreting the results. There should therefore be a trade-off between the optimum number of topics and the size of the documents 'corpus being analysed. This choice cannot be relied on the automatism of the method alone, but must be appropriately guided by knowledge of the subject on which the scientist is working.

Materials and methods

The statistical analysis was conducted with R 4.0.5 (R Core Team 2021), using the library 'tm' (Feinerer et al. 2008) and 'topicmodels' (Grün and Hornik 2011).

Consistent with previous articles (Contiero et al. 2019; Nalon et al. 2021) a literature search protocol was performed on Scopus®, the bibliographic and citation database of Elsevier©, to assess research trends on reindeer pastoralism. After a screening procedure described in Holand et al. (2024), a corpus of 2875 abstracts dealing with reindeer pastoralism literature was retained for the subsequent analyses. The corpus of abstracts was transformed into a numerical representation, the documents to terms matrix (DTM) on which it is possible to operate mathematically. A text mining and topic analysis were conducted to extract the latent structure of meaningful themes from the collection of textual information, represented by the articles' abstracts. In our study, we utilised a topic modelling approach with the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003) to uncover the underlying structure of meaningful themes from a collection of abstracts. The algorithm identifies a latent structure of thematic topics based on the co-occurrence of words within the documents. This technique assumes each document is a mix of topics, and each topic is a mix of words. At the end of the process, each document is assigned to a topic with the highest probability, and each word is ranked inside the topic according to its probability. Based on these attributions, the groups of articles were assigned the topics names (labels) by the experts involved in the analysis. In a following collective time, the labels were subjected to an in-depth discussion afterward the authors team reached a consensus with regards to the most appropriate final labels for each topic. As final step, a hierarchical cluster analysis approach was carried out to better understand the proximity of the resulted topics (Wang et al. 2016).

The algorithm identified the groups and the related documents and terms distribution but it did not provide the optimal number of topics to be considered. Four approaches were therefore performed to identify the suitable number of topics that better summarise the information of the reindeer pastoralism data corpus, bearing in mind, that the reduction of qualitative content through a mathematical approach is always limiting.

1. The perplexity index measures the uncertainty in a predicted model and it is assumed that a lower perplexity score indicates better performance (Griffiths and Steyvers 2004). This machine learning approach allows to test the adequacy of a model, developed in a train set, measuring its performance on a hold-out set. In our study, the

DTM was randomly split into two parts: the former including the 80% of the documents was used as train set and the latter as test set (hold-out set). For different numbers of topics (from 2 to 20) the algorithm was fitted on the train set. Using the results obtained in the train set, the perplexity index was calculated both for the train and the hold-out sets (using the function perplexity contained in topicmodels package). The optimal number of topics corresponds to the minimum value of the function.

2. The harmonic mean of the log-likelihood of a set of samples generated by the Gibbs sampler (Griffiths and Steyvers 2004) is an index to assess the optimal number of topics, considering that in an iterative process comparing a different number of topics, a higher value is better. The iterative process was developed from 2 to 20 topics. The index was calculated using the logPwzT function of topicmodels package (Ponweiser 2012).
3. The number of articles assigned to a given topic can be a further empirical criterion for choosing the optimal number of topics. It can be assumed that a too low percentage of articles per topic hampers a clear interpretability of the total corpus of papers. Therefore, analyses with a too high number of topics (and a consequent low percentage of assigned articles per each topic) in relation to the size of the corpus may be counterproductive.
4. The hierarchical cluster analysis carried out using different number of topics produces the estimation of the within-classes and between class explained variance that could be considered as indexes of the goodness of fit of the simulation. The goal of a partitioning clustering algorithms is to split the dataset into groups such that the objects in the same cluster are similar as much as possible (compactness) and the objects in different clusters are highly distinct (separation). A lower within-cluster variation is an indicator of a good compactness of the objects inside the groups, whereas a higher between-cluster variation is an indicator of a well clusters' separation. The hierarchical clustering analysis (Wang et al. 2016; Nalon et al. 2021) was calculated to group topics based on the topic-word matrix (topic by the rows and words by the columns). This is a binary matrix with 1/0 to indicate the presence of a word (along columns) in a given topic (along rows). The distance among topics was calculated based on the Jaccard distance and the average linkage method was applied with an agglomerative clustering algorithm to generate the cluster dendrogram. The analysis was conducted in XLSTAT (Lumivero 2024, XLSTAT statistical and data solution. https://www.xlstat.com/en).

Results and discussion

The choice of the optimal number of topics that best explain the main information from a large data corpus does not always provide clear and unambiguous results (Wang et al. 2016). As the 'ideal' number is in general not known, the perplexity index and the log-likelihood harmonic mean were calculated for documents corpus of reindeer pastoralism assuming an increasing number of topics from 2 to 20 (Figure 1).

Since both indexes did not give precise indications regarding the choice of the optimal number of topics to be considered (no detectable minimum or maximum for perplexity and harmonic mean, respectively), four different simulations were carried out with 8, 10, 12, and 20 topics. The 20 most probable words of each topic obtained from the four simulations and their interpretation (expressed as label and short label) by the experts are listened in Table 1. The choice of the label of each topic was based either on the 20 most probable words or on the analysis of the abstracts of the assigned articles.

The procedure highlighted 4 topics that in all the four simulations remained constants and quite distinct from the others. This is the case of the 'Socio-cultural' aspects, 'Nutrition and meat production', 'Herd productivity- population dynamic', and 'Archaeology' (Table 1). These represent established research strands that remain ever-present regardless of the number of topics you choose. On the other hand, the simulations with an increasing number of topics enabled to clearly and distinctly disclose new arguments that were previously fully or partially hidden. Considering the Figure 2 which reports the contingency tables obtained crossing the 10 topics simulation (by rows) with a) 8-topics (by columns and b) 12-topics (columns) simulations, it can be noticed that in the 8-topics simulation different issues were lumped together within the same group. For example, articles related to radioactive contamination caused by the Chernobyl accident, which were classified within the 'Feeding and Food' group by 8-topics simulation (Figure 2a), were then assigned to health-related articles ('Health and diseases') in the 10-topics simulation. Radioactivity contaminated lichens entered the reindeer feeding chain and consequently affected the

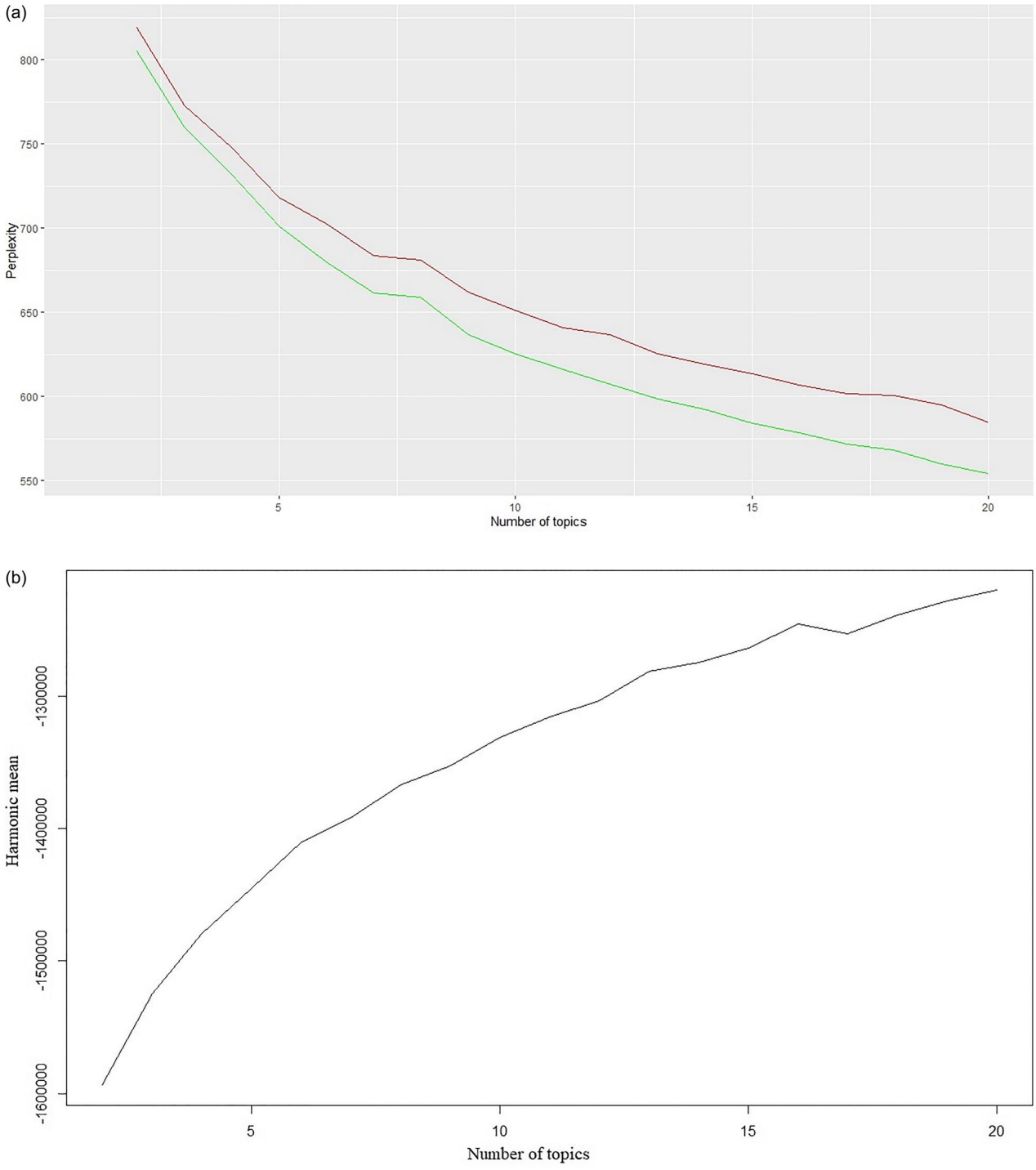

Figure 1. Two indexes for a sequence of different number of topics (from 2 to 20): (a) perplexity index of training (green) and hold-out (red) datasets; (b) harmonic mean of the log-likelihood.

health of the animals and the safety of the meat products they supply. Indeed, the passage from 8- to 10-topics simulation brought out issues that were not previously detected: i.e. the group 'Health and diseases' both for animal and herders.

The crossing between the 12-topics simulation (by column) and the 10-topics one (by row; Figure 2(b) did not lead to the identification of further relevant arguments for interpreting the data corpus. A single additional information regarded the partition of the socio-cultural topic into two topics regarding two ethnic groups traditionally dedicated to reindeer herding: the former embracing the socio-cultural aspects of the Sami population and the latter focusing on traditional and economic aspects of Russian nomads (mainly Nenets and other Russian ethnic groups). As might be

Table 1. The 20 most probable words[a], extended and short label, number of articles and percentage for the four simulations (8-, 10-, 12-, and 20-topics).

Simulation	Topics	20 Most probable words	Label	Short label	Documents
8 topics	1	develop people herder culture indigen tradit natur Sami research local social econom state articl practice region knowledge land paper resource	Social, cultural and economic aspects	Socio-cultural	685 (24%)
	2	graze speci arctic plant effect veget increas tundra soil abund ecosystem community* product chang result herbivor studi differ high signific	Vegetation interactions	Grazing and vegetation	287 (10%)
	3	site hunt human remain part bone antler also earli late time evid period materi provid present include anim date well	Archaeology	Archaeology	351 (12%)
	4	winter increas femal popul season year time male rate calv bodi summer herd rangif condit effect size variat mass reproduct	Herd productivity and population dynamic	Herd productivity	352 (12%)
	5	popul differ group studi among wild result analysi show method three domest number data base genet identify analys structur reveal	Large herbivore interactions	Herbivore interaction	227 (8%)
	6	level anim* sampl* concentr* food deer meat diet found speci valu signific higher test respect collect determin compar total content	Nutrition and meat production	Feeding and food	490 (17%)
	7	model habitat rang select predate manag spatial landscap speci within distribut movement resource ecolog activ conserv area human respons behaviour	Landscape ecology	Space use	267 (9%)
	8	area chang lichen forest climat studi northern factor pastur cover snow impact also import condit Finland region main mountain Norway	Climate change impact	Climate change	216 (8%)
10 topics	1	chang arctic climat condit environment impact increas snow temperature factor adapt result environ respons warm import variabl influenc pattern interact	Climate change impact	Climate change	167 (6%)
	2	lichen graze forest plant veget effect soil speci cover tundra increas community differ abund area herbivor result ecosystem product shrub	Vegetation interactions	Grazing and vegetation	318 (11%)
	3	level food concentr sampl valu meat diet higher active signific high product found determin total measure content respect compar collect	Nutrition and meat production	Feeding and food	365 (13%)
	4	area northern studi data Norway part region pastur Sweden present also Finland year number period mountain main time century sinc	Pasture utilisation in Fennoscandia	Pasture in Fennoscandia	115 (4%)
	5	people develop herder indigen tradit culture research natur social econom local articl practice state paper knowledge manag industry Russian discuss	Social, cultural and economic aspects	Socio-cultural	619 (22%)
	6	winter femal increas season male summer rate calv popul bodi year rangif time herd effect mass reproduct size adult growth	Herd productivity and population dynamic	Herd productivity	299 (10%)
	7	model rang habitat select manag predate spatial resource landscap movement conserve larg behaviour distribut area prey scale within estim line	Landscape ecology	Space use	234 (8%)
	8	site human hunt remain antler bone late larg date also materi locat evid include cave earli europ point analysi provid	Archaeology	Archaeology	312 (11%)
	9	popul speci differ deer wild domest genet herd rangif found moos breed indic analysi structure number show island sheep analys	Large herbivore interactions	Herbivore interaction	231 (8%)
	10	anim group studi among Sami differ method test result relat compar infect diseas report parasite identify case specif investing signific	Health and diseases	Health and diseases	215 (7%)
12 topics	1	area forest northern Norway number studi import also year mountain Finland period part main howev present major sever sinc Sweden	Pasture utilisation in Fennoscandia	Pasture in Fennoscandia	143 (5%)
	2	manag model system resourc ecolog natur provid assess conserv develop land estim approach includ impact potenti harvest current landscap need	Natural resource management	Resources management	168 (6%)
	3	lichen graze plant veget speci effect soil increas communiti tundra cover herbivor product abund shrub biomass composit high domin ecosystem	Vegetation interactions	Grazing and vegetation	284 (10%)
	4	differ studi data result group show analysi method three observ analys relat base type also compar structur investig present inform	Modelling studies	Modelling studies	93 (3%)
	5	herder Sami cultur local social discuss paper relat knowledg research among focus practic adapt human articl govern polici right Saami	Sami's social and cultural aspects	Sami socio-cultural	359 (12%)

(*continued*)

Table 1. Continued.

Simulation	Topics	20 Most probable words	Label	Short label	Documents
	6	region people develop indigen tradit north natur state econom life industri Russian breed live activ tundra nomad Nenet territori cultur	Russian herders' social, cultural and economic aspects	Russian herders	273 (9%)
	7	rang habitat select predat spatial movement distribut activ within human rangif larg area scale popul prey behaviour disturb line lynx	Landscape ecology and predations	Space use	201 (7%)
	8	chang climat arctic increas condit snow temperatur environment popul factor impact effect affect time warm influenc environ associ dynam respons	Climate change impact	Climate change	153 (5%)
	9	level food concentr sampl anim diet meat higher signific valu high activ content collect determin compar respect milk total acid	Nutrition and meat production	Feeding and food	329 (11%)
	10	site remain bone hunt antler earli late larg date human materi evid suggest cave locat upper period part includ Europ	Archaeology	Archaeology	298 (10%)
	11	popul speci deer anim wild genet herd domest rangif sampl found infect among report test diseas moos parasit detect isol	Health diseases wild and domestic herbivores interactions	Health and diseases	316 (11%)
	12	femal winter male calv bodi rate increas season year summer herd popul size mass reproduct time adult variat effect growth	Herd productivity and applied population dynamic	Herd productivity	258 (9%)
20 topics	1	graze plant veget soil tundra communiti herbivor site shrub biomass effect speci product ecosystem abund plot moss nutrient cover domin	Vegetation interactions	Grazing and vegetation	202 (7%)
	2	speci three type identifi differ isol analys sequenc present number analysi show four reveal rumen found gene previous repres within	Rumen microbiota	Rumen microbiota	101 (4%)
	3	chang climat arctic environment temperatur condit impact environ region warm factor event adapt recent global affect weather respons result extrem	Climate change impact	Climate change	137 (5%)
	4	time activ period year season occur pattern follow first migrat earli record cycl continu maximum sinc throughout least peak earlier	Seasonal dynamic	Seasonal dynamic	56 (2%)
	5	model data base method studi analysi result estim field assess inform combin provid also process includ approach appli monitor compon	Scientific terminology	Scientific terminology	106 (4%)
	6	herder cultur social local research paper knowledg discuss focus practic govern human relat right saami process understand adapt role context	Sami's social and cultural aspects	Sami socio-cultural	277 (10%)
	7	concentr food meat sampl level content milk acid diet product valu intak dose collect higher determin element contamin total protein	Nutrition and meat production	Feeding and food	255 (9%)
	8	peopl tradit indigen develop region Russian north state life econom articl* nomad nenet industri territori natur cultur author Russia economi	Russian herders' social, cultural and economic aspects	Russian herders	241 (8%)
	9	site bone remain hunt human cave date late materi antler upper hors archaeolog part locat evid assemblag also suggest earli	Archaeology	Archaeology	275 (10%)
	10	area forest northern mountain Norway Finland studi tree boreal southern sever establish common part finnish centuri sweden fire stand south	Forest pasture in Fennoscandia	Forest in Fennoscandia	97 (3%)
	11	habitat select spatial rang movement human landscap scale within line disturb respons distribut distanc pattern behaviour resourc featur individu across	Landscape ecology	Space use	130 (5%)
	12	popul herd wild domest genet island structur breed rangif size north variat geograph alaska origin among distribut divers rang region	Large herbivores' interactions	Herbivores interaction	105 (4%)
	13	increas effect rate growth reduc declin densiti popul caus affect negat howev decreas high dynam depend suggest control limit year	Population dynamic	Population dynamic	54 (2%)
	14	manag system resourc land natur conserv develop protect harvest product sustain plan ecolog impact potenti current project econom conflict includ	Land management and protection	Land management	134 (5%)
	15	anim deer sampl test infect parasit rangif diseas host moos detect examin cervid sheep serum report specif collect preval case	Health and diseases and wild and domestic herbivores interactions	Health and disease	192 (7%)
	16	larg predat speci import prey bear small lynx area main abund wolf number limit mammal wolverin wolv found ungul kill	Carnivours predation	Predation	103 (4%)
	17	differ signific studi result compar show level observ relat indic higher found valu high lower similar well mean correl investig	Statistical terminology	Statistical terminology	44 (2%)

(*continued*)

Table 1. Continued.

Simulation	Topics	20 Most probable words	Label	Short label	Documents
	18	femal male calv bodi mass reproduct adult weight size individu success surviv ratio antler calf rangif variat length condit young	Reproduction biology	Reproduction biology	173 (6%)
	19	lichen winter summer forag snow rang feed pastur condit svalbard cover high energi qualiti season ground measur import avail compar	Summer and winter feeding	Pasture and forage	104 (4%)
	20	group sami among studi live northern work herder health sweden associ swedish languag risk tourism part factor women mortal investig	Swedish Sami's health	Health Swedish Sami	89 (3%)

[a]Many words were reduced to their root form by the algorithm (Porter 1980).

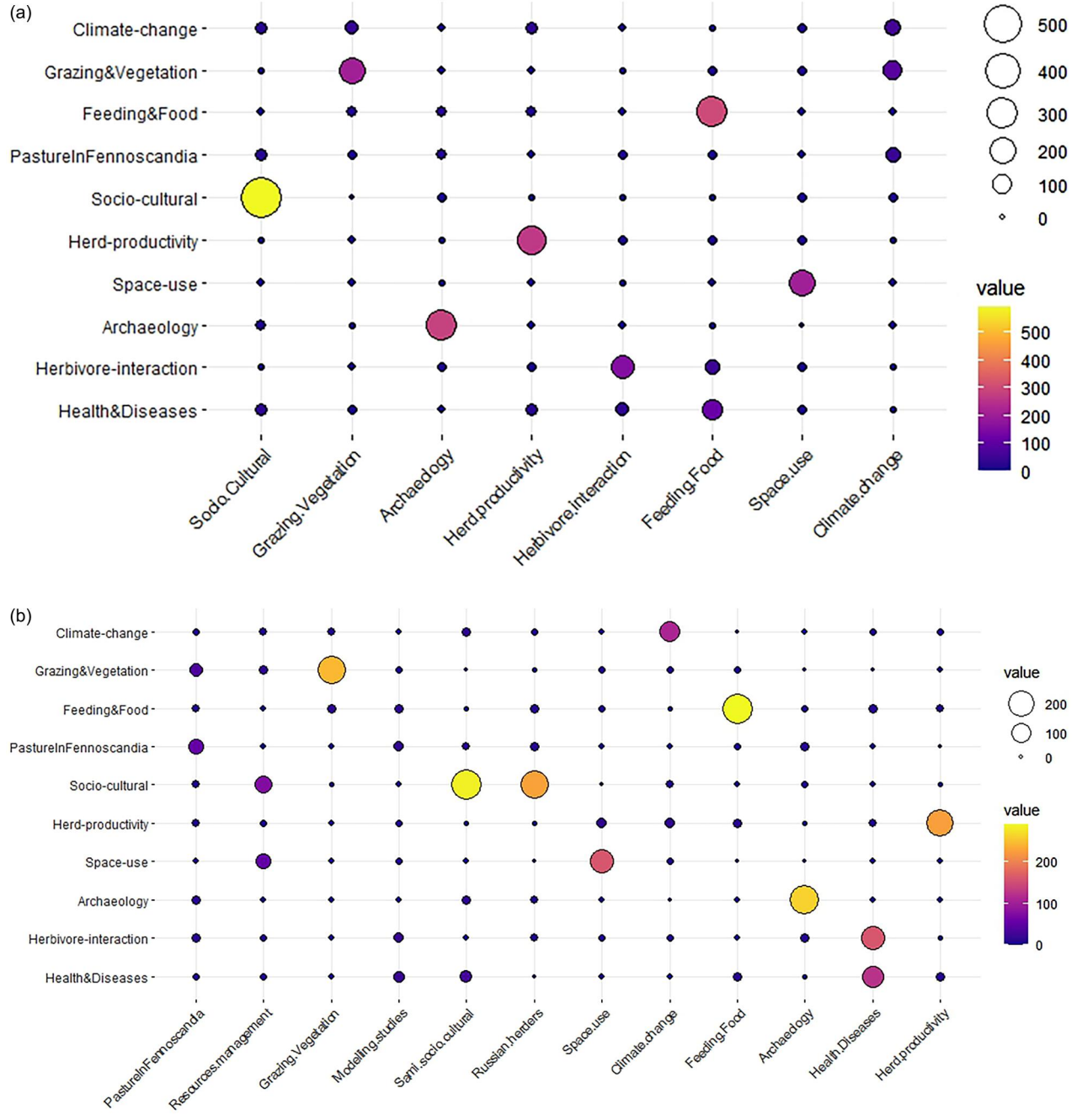

Figure 2. Balloon plots of the contingency tables obtained crossing the 10-topics simulation (rows) with a) 8-topics (columns) and b) 12-topics (columns) simulation. Each dot represents the number of articles shared by the corresponding topics in the two simulations. The size and colour of each dot reflects the relative magnitude of the corresponding component.

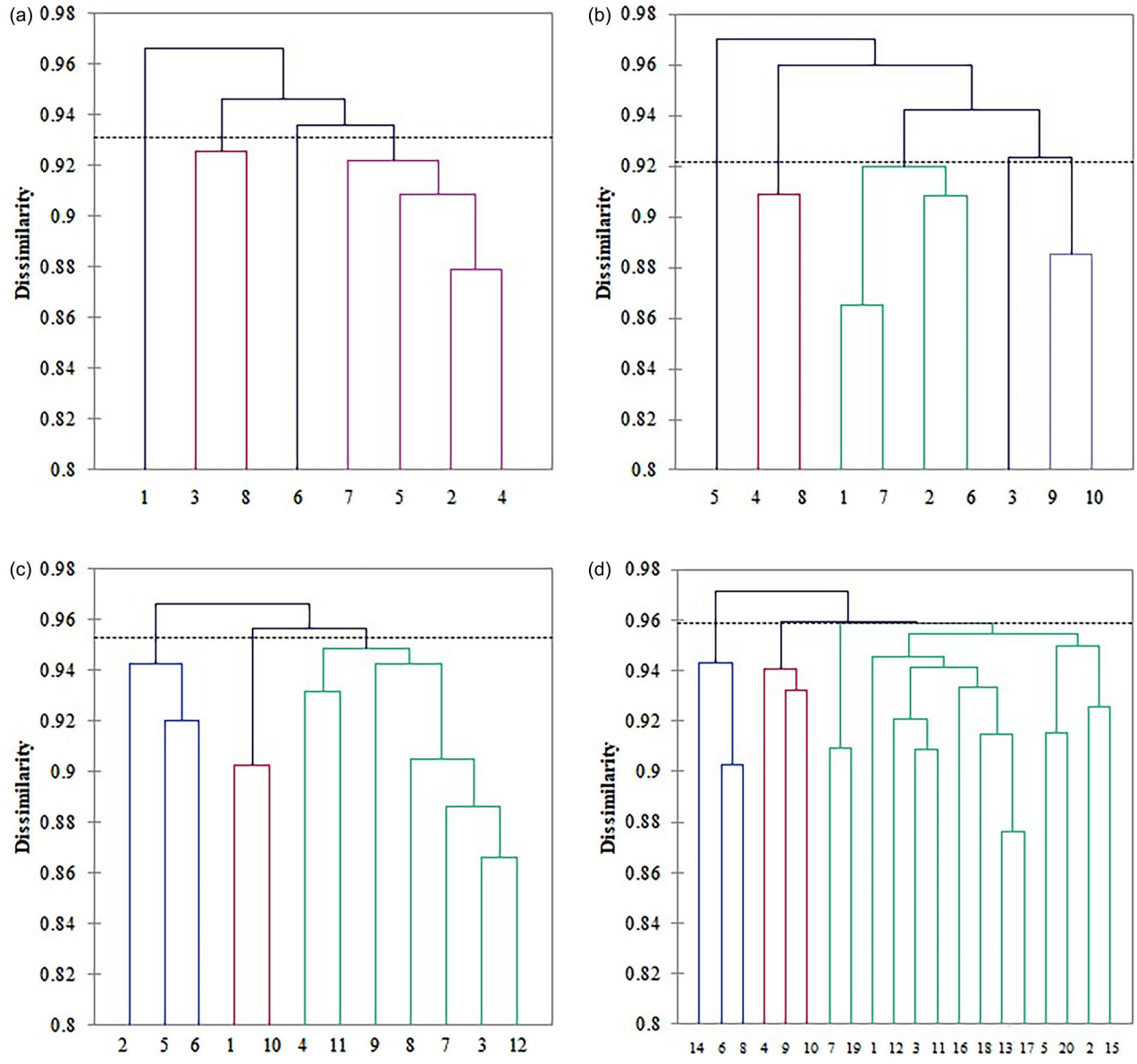

Figure 3. Dendrograms of the hierarchical cluster analysis of the topics based on dissimilarity matrix of the terms contained in the topics for the 4 simulations: (a) 8-topics; (b) 10-topics; (c) 12-topics; (d) 20-topics. The dotted line represents the tree cut at the maximum entropy index.

expected, the 20-topics simulation brought to light a few arguments that had never been highlighted in previous analyses (as 'Predation'), but it mainly further partitioned into separate groups of arguments that were already been identified. On the other hand, this simulation tended to identify topics with very general meanings, such as 'Scientific terminology', 'Statistical terminology', or too specific such 'Health of Swedish Sami' (Table 1). Simulations with an increasing number of topics lead to a progressive fragmentation of the data corpus creating small aggregations of articles in percentage terms. This trend can be visually appreciated through the reduction of the size of the balloons moving from Figure 2a to Figure 2b. These findings suggest that a percentage of articles per topic lower than 5% is not so relevant for a deep interpretation of a 3000 articles data corpus (as the one under consideration). Based on these outcomes, the 10-topic simulation should be enough detailed to allow a comprehensive interpretation of the reindeer pastoralism literature used as case study.

The hierarchical clusterisation of the dissimilarity matrix between terms belonging to different topics can be useful for understanding both how topics are grouped together (and therefore in some sense close) and what the best choice among the number of topics might be (Figure 3). The within class explained variance values for the four simulations

were: 95, 92, 96, and 98% for 8-, 10-, 12-, and 20-topics simulations, respectively. Quite surprisingly, it was not monotonically increasing as found by Nalon et al. (2021). The fact that the 10-topics simulation provides a lower explained variance value than the others is probably due to the higher number of clusters that were created with this simulation (5 clusters *vs.* 3/4 of the others). This implies that the compactness within groups is greater and thus clustering works better. Therefore, the 10-topics simulation appears to be the most functional solution for an in-depth interpretation of the data corpus as it keeps apart relevant topics that would otherwise be collapsed into larger aggregations (Holand et al. 2024).

Conclusions

Although text mining techniques appear to be objective mathematical approaches, the method is not without its weaknesses. Indeed, all steps in the process must be scrutinised carefully and the results have to be deeply analysed and put into the right context to generate informative and comprehensive knowledge. The results of this study showed that there is no an ultimate method for achieving a balanced compromise between the detailed resolution of a given corpus of documents and the perplexity index when using a text mining analysis. The appropriate number of topics to consider should take into account the size of the data corpus. Excessive fragmentation of the data corpus into small document aggregations encourages the emergence of topics devoid of any technical or practical meaning, solely as a result of the unsupervised iterative process. Conversely, if subsequent clusterisation yields high values of explained variability within groups, this index can serve as an indicator for selecting the optimal number of groups.

Ethical approval

Not required.

Disclosure statement

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

Funding

This research received no external funding.

ORCID

Barbara Contiero http://orcid.org/0000-0002-6586-5035
Øystein Holand http://orcid.org/0000-0002-8830-4310
Giulio Cozzi http://orcid.org/0000-0003-0408-1082

Data availability statement

Dataset available on request from the authors.

References

Adamaκopoulou C, Benedetti B, Zappaterra M, Felici M, Masebo NT, Previti A, Passantino A, Padalino B. 2023. Cats' and dogs' welfare: text mining and topics modelling analysis of the scientific literature. Front Vet Sci. 10:1268821. doi: 10.3389/fvets.2023.1268821.

Antons D, Grünwald E, Cichy P, Salge TO. 2020. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. R&D Manage. 50(3):329–351. doi: 10.1111/radm.12408.

Benedetti B, Felici M, Nanni Costa L, Padalino B. 2023. A review of horse welfare literature from 1980 to 2023 with a text mining and topic analysis approach. Ital J Anim Sci. 22(1):1095–1109. doi: 10.1080/1828051X.2023.2271038.

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. J Mach Learn Res. 3:993–1022. doi: 10.5555/944919.944937.

Bornmann L, Mutz R. 2015. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. Assoc Info Sci Tech. 66(11):2215–2222. doi: 10.1002/asi.23329.

Brscic M, Contiero B, Schianchi A, Marogna C. 2021. Challenging suicide, burnout, and depression among veterinary practitioners and students: text mining and topics modelling analysis of the scientific literature. BMC Vet Res. 17(1):294. doi: 10.1186/s12917-021-03000-x.

Contiero B, Cozzi G, Karpf L, Gottardo F. 2019. Pain in pig production: text mining analysis of the scientific literature. J Agric Environ Ethics. 32(3):401–412. doi: 10.1007/s10806-019-09781-4.

Feinerer I, Hornik K, Meyer D. 2008. Text mining infrastructure in R. J Stat Soft. 25(5):1–54. doi: 10.18637/jss.v025.i05.

Griffiths TL, Steyvers M. 2004. Finding scientific topics. Proc Natl Acad Sci USA. 101(Suppl 1):5228–5235. doi: 10.1073/pnas.0307752101.

Grün B, Hornik K. 2011. topicmodels: an R package for fitting topic models. J Stat Soft. 40(13):1–30. doi: 10.18637/jss.v040.i13.

Holand Ø, Contiero B, Næss MW, Cozzi G. 2024. "The time they are a-changin' – research trend and perspectives of reindeer pastoralism – A review using text mining and topic modelling. Land Use Policy. 136:106976. doi: 10.1016/j.landusepol.2023.106976.

Jain AK. 2010. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 31(8):651–666. doi: 10.1016/j.patrec.2009.09.011.

Masebo NT, Zappaterra M, Felici M, Benedetti B, Padalino B. 2023. Dromedary camel's welfare: literature from 1980 to 2023 with a text mining and topic analysis approach.

Front Vet Sci. 10:1277512. doi: 10.3389/fvets.2023.1277512.

Nalon E, Contiero B, Gottardo F, Cozzi G. 2021. The welfare of beef cattle in the scientific literature from 1990 to 2019: a text mining approach. Front Vet Sci. 7:588749. doi: 10.3389/fvets.2020.588749.

O'Connor A, Sargeant J. 2015. Research synthesis in veterinary science: narrative reviews, systematic reviews and meta-analysis. Vet J. 206(3):261–267. doi: 10.1016/j.tvjl.2015.08.025.

Pape R, Löffler J. 2012. Climate change, land use conflicts, predation and ecological degradation as challenges for reindeer husbandry in Northern Europe: what do we really know after half a century of research? Ambio. 41(5):421–434. doi: 10.1007/s13280-012-0257-6.

Park K, Kremer G. 2017. Text mining-based categorization and user perspective analysis of environmental sustainability indicators for manufacturing and service systems. Ecol Indic. 72:803–820. doi: 10.1016/j.ecolind.2016.08.027.

Ponweiser M. 2012. Latent Dirichlet Allocation in R [theses]. WU Vienna University of Economics and Business. Institute for Statistics and Mathematics, No. 2. doi: 10.57938/533618e5-dcd9-4c8f-913a-2339fa145c71.

Porter MF. 1980. An algorithm for suffix stripping. Program. 14(3):130–137. doi: 10.1108/eb046814.

R Core Team 2021. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Trapanese L, Petrocchi Jasinski F, Bifulco G, Pasquino N, Bernabucci U, Salzano A. 2024. Buffalo welfare: a literature review from 1992 to 2023 with a text mining and topic analysis approach. Ital J Anim Sci. 23(1):570–584. doi: 10.1080/1828051X.2024.2333813.

Tucciarone I, Secci G, Contiero B, Parisi G. 2024. Sustainable aquaculture over the last 30 years: an analysis of the scientific literature by the text mining approach. Rev Aquacult. 1–13. doi: 10.1111/raq.12950.

Wang S-H, Ding Y, Zhao W, Huang Y-H, Perkins R, Zou W, Chen JJ. 2016. Text mining for identifying topics in the literatures about adolescent substance use and depression. BMC Public Health. 16(1):279. doi: 10.1186/s12889-016-2932-1.

Zuliani A, Contiero B, Schneider MK, Arsenos G, Bernués A, Dovc P, Gauly M, Holand Ø, Martin B, Morgan-Davies C, et al. 2021. Topics and trends in mountain livestock farming research: a text mining approach. Animal. 15(1):100058. doi: 10.1016/j.animal.2020.100058.