

GOOD AND BAD SOCIOLOGY: DOES TOPIC MODELLING MAKE A DIFFERENCE?

MARIUSZ BARANOWSKI¹ & PIOTR CICHOCKI²

¹ Adam Mickiewicz University in Poznań, Szamarzewskiego 89 C, 60-568 Poznań, Poland. ORCID: 0000-0001-6755-9368, Email: <u>mariusz.baranowski@amu.edu.pl</u>

² Adam Mickiewicz University in Poznań, Szamarzewskiego 89 C, 60-568 Poznań, Poland. ORCID: 0000-0002-6501-9082, Email: <u>piotr.cichocki@amu.edu.pl</u>

ABSTRACT: The changing social reality, which is increasingly digitally networked, requires new research methods capable of analysing large bodies of data (including textual data). This development poses a challenge for sociology, whose ambition is primarily to describe and explain social reality. As traditional sociological research methods focus on analysing relatively small data, the existential challenge of today involves the need to embrace new methods and techniques, which enable valuable insights into big volumes of data at speed. One such emerging area of investigation involves the application of Natural Language Processing and Machine-Learning to text mining, which allows for swift analyses of vast bodies of textual content. The paper's main aim is to probe whether such a novel approach, namely, topic modelling based on Latent Dirichlet Allocation (LDA) algorithm, can find meaningful applications within sociology and whether its adaptation makes sociology perform its tasks better. In order to outline the context of the applicability of LDA in the social sciences and humanities, an analysis of abstracts of articles published in journals indexed in Elsevier's Scopus database on topic modelling was conducted. This study, based on 1,149 abstracts, showed not only the diversity of topics undertaken by researchers but helped to answer the question of whether sociology using topic modelling is "good" sociology in the sense that it provides opportunities for exploration of topic areas and data that would not otherwise be undertaken.

KEYWORDS: unsupervised text analysis, LDA, topic modelling, sociological methods, big data sociology

INTRODUCTION

In taking up the good and bad sociology theme, it is essential to remember that "sociology has always been a divided and controversial field" (Wilterdink 2012: 2). The same is true of the research methods used in sociological sub-disciplines, which further reinforce the conflicts and divisions within this social science (cf. Gouldner 1976). The principal juxtaposition arose between the quantitative approaches, usually based on surveys, and the qualitative domain, which mostly involved interpretative studies of textual data, e.g., interview transcripts, personal documents, media discourse. The ready availability of large amounts of digital information and the rise of powerful computational methods of analysing undermine those established distinctions and pose a significant challenge for sociology in a technologically networked society (Baranowski 2021; Selwyn 2015). Sociological methods require serious re-thinking, and there is a pressing need for developing new methods or adapting those already developed in other disciplines. Otherwise, sociology would remain stuck with a twentieth-century tool-set and risk sliding towards irrelevance and obscurity (Adorjan & Kelly 2021; Baranowski & Mroczkowska 2021).

One of the recently developed approaches allowing large-scale and rapid information extraction out of vast text bodies is topic modelling. It should be noted-following Hannigan et al. (2019: 589)-that "borrowed from computer science, this method involves using algorithms to analyse a corpus (a set of textual documents) to generate a representation of the latent topics discussed therein (Mohr & Bogdanov 2013; Schmiedel, Müller, & vom Brocke 2018)". Topic modelling, most commonly based on Latent Dirichlet Allocation (LDA) (Silge & Robinson 2017: 89-108), represents one of the recent advances in Natural Language Processing, which bring about massively enhanced opportunities for using content analysis in the context of sociological research. In general, it involves a fundamental transition from the survey mindset-extrapolation from small samples of carefully curated and structured data—to the Big Data mindset where large volumes of loosely organised information are processed at speed to discern the signals from the overall noise. On the other hand, while being a statistical approach, topic modelling defies the traditional juxtaposition of quality vs quantity, as it produces formal quantitative insights into the qualitative domain of meaning.

While quantitative content analysis existed within the classical paradigm of social sciences (Weber 1990), it was notably limited in its applications due to its low scalability. Consequently, the study of textual data has traditionally become a domain dominated by qualitative approaches, with quantitative accounts subsisting at the margins. Text-focused machine learning techniques remove the scalability limitations, as algorithms can read previously unimaginable volumes of text and handle highly complex coding schemes; their "reading" is also controlled by explicit parameter settings and, therefore, replicable. In particular, topic modelling allows for identifying the thematic structure of a large corpus of documents, which roughly resembles the highlighting technique of classical content analysis in that it marks every document—or some of its smaller constituent parts—with the probability of belonging to each of the topics identified in the overall corpus of documents. Apart from the easy scalability, topic modelling also allows for multiple iterations and, unlike its classical counterpart, is not bound by a coding scheme fixed before the commencement of analysis. At first glance, topic modelling also seems to reduce the arbitrary impact of human subjectivity in that it does away with human coders; however, even though the analysis is explicitly specified in the form of replicable computer code, the decisions made by the code-writing analyst shape the model outputs in powerful and often not entirely predictable ways.

This paper demonstrates how topic analysis can be implemented into sociological inquiry making sociology "good", i.e. at least better off than with its current inventory of methods. Based on an LDA analysis performed on 1149 abstracts of academic texts mentioning topic modelling, we (i) discuss the application of LDA in the context of traditional methods of content analysis, (ii) present basic insights into the thematic structure of articles using topic modelling, (iii) conclude with recommendations concerning future use and research opportunities associated with computational approaches to the analysis of textual data.

LITERATURE REVIEW

The need for systematic content analysis yielding quantifiable results was recognised at the outset of modern social sciences—even before the first world war, Max Weber put forward a proposal for exhaustive press monitoring to measure the "cultural temperature" of society (Lazarsfeld & Oberschall 1965). While Weber's ideas could not be matched by any existing methodological tools and research infrastructures, pioneering research on the press's discourse was empirically taken up by the next generation of researchers, most notably perhaps by Harold Lasswell (1927). However, the formulation of the classical paradigm of content analysis is typically associated with Bernald Berleson's stipulation that it amounts to a "research technique for the objective, systematic and quantitative description of the manifest content of communication" (1952: 18). According to Berleson, content analysis was supposed to serve the following five aims:

- 1. Describing the substance characteristics of message content
- 2. Describing the form characteristics of message content
- 3. Making inferences to producers of content
- 4. Making inferences to audiences of content
- 5. Predicting the effects of content on audiences

In the narrow sense, the classical content analysis boiled down to the first two of those aims as it matured into a set of techniques geared towards the systematic classification of communications allowing for exploring their content and formal characteristics. In terms of the expected structure of the research process, the quantitative content analysis came to be conceived as a survey with documents standing in for respondents. Due to its limited significance in social sciences, the field also did not benefit from much innovation. The postulated sequence of steps required to conduct a content analysis remained fixed over the decades: (1) formulating the research problem, (2) selecting a sample of content, (3) determining the units of analysis, (4) specifying the coding scheme, (5) coding and (6) statistical analysis of the coding output (Mayntz, Holm, & Hübner 1976).

The classical approach suffers from several substantial limitations. Firstly, the necessity of sampling due to the practical impossibility of reading all content, which can arguably be performed better than in human surveys as documents have higher response rates, but still brings about some sampling error. Secondly, the essentially aprioristic nature of the coding scheme-although some free reading and pilot research is involved in its specification, it can only be based on a very limited selection of documents. Thirdly, the need for training coders and maintaining intercoder reliability-the true Achilles heel of the whole process as it limits both coding complexity and empirical scalability. In order for the coders to be reliably consistent in their judgements, they need to be well trained based on a uniform and unambiguous set of coding variables and instructions, which makes it necessary to keep them short and straightforward. Furthermore, as increasing the number of coders strains the coherence of the coder group and necessitates ambitious quality-control schemes, the analysis does not scale well and is hardly replicable as any future re-implementation requires re-training of the coders. Fourthly, the statistical analysis of results bound by the coding data it receives—in theory, it could lead to changes in the coding scheme and instructions, but this would require a costly re-run of most of the content-analytic process.

Being aware of the limitations of the classical approach and the advantages of topic modelling, let us quote the point of view of Monica Lee and John Levi Martin, who referred explicitly to good and bad sociologists. Quite provocatively, they stated that:

When it comes to formal analyses, we might say that bad sociologists code, and good sociologists count. The reason is that the former disguises the interpretation and moves it backstage, while the latter delays the interpretation, and then presents the reader with the same data on which to make an interpretation that the researcher herself uses. Even more, the precise outlines of the impoverishment procedure is explicit and easily communicated to others for their critique. And it is this fundamentally shared and open characteristic that we think is most laudable about the formal approach. (Lee & Martin 2015: 24)

We take the above statement instead as forcing a discussion and a critical examination of the status of methods used in sociology, since, as Juho Pääkkönen and Petri Ylikoski (2021: 1469) have pointed out, "it remains unclear how unsupervised methods can, in fact, support interpretative work and in what sense this could be said to make interpretation more objective". Unsupervised topic modelling, which we treat as an exploratory technique, does not aim at superseding the traditional approaches to content analysis; it does, however, constitute an interesting complement to them. In terms of objectivity, it clearly does away with substantial amounts of subjectivity by way of eliminating the human readers, whose interpretative decisions prove difficult to account, especially in the contexts of qualitative content analysis. On the other hand, topic modelling relies on a number of arbitrary choices, e.g. setting the number of desired topics or determining the values of hyper-parameters, and its results are also highly dependent on the procedures applied for text cleaning and pre-processing. Although, all those decisions are made explicit in the code script, and therefore fully replicable; furthermore, a variety of metrics exist which guide the analyst towards better choices. Yet, all this provides a framework for managing subjectivity, rather than a solution eliminating it.

METHODOLOGY

Our LDA analysis (Blei, Ng, & Jordan 2003) was performed on abstracts of articles containing the phrases "topic modelling", "topic modeling" and "topic model" in their abstract or title, which were downloaded from Scopus for the period 2000-2020. The database comprises 1,149 individual records containing the abstract and publication meta-data, e.g. citation count or author affiliation; however, the meta-data is not made available to the LDA algorithm. It can be merged with the LDA outputs at a later stage. The analysis proceeded in three distinct stages: (i) data cleaning, (ii) modelling and parameter setting, (iii) model exploration. The analysis was performed in R (R Core Team 2021), within the family of libraries associated with tidyverse() and tidytext() packages.

Since LDA requires a document-term matrix, extensive data-cleaning and preprocessing are needed for the algorithm to work correctly. Crucially, it must be noted that LDA-techniques belong to the general approach known as "bag-of-words", which amounts to treating every document as an unordered list of tokens (usually but not necessarily individual words or common phrases). Topics are defined based on the probability distribution of specific tokens within a given vocabulary.

The extent of necessary data cleaning varies depending on several factors-principally, however, on the shape and quality of the text input. Crucially, however, the cleaning and reshaping code usually remains re-usable across different projects with only minor tweaks required. Therefore, it is easy to scale up and repurpose once the LDA workflow is set up. Firstly, the data needs to be imported and initially cleaned (for instance, every Scopus abstract contains a copyright statement at the end, which needs to be erased as it leads the algorithm to seek topics based on the names of the major publishing houses). Secondly, Natural Language Processing needs to happen, which in our case involves using the Spacy library for tokenisation, POS-recognition and nounphrase extraction. Note that after Spacy, only noun phrases are retained (based on the assumption that nominal elements of the argument structure carry the relevant information). Following the extraction of noun phrases, some additional housekeeping commences: (1) rough spell-check using Hunspell, principally aimed at unifying British and American spellings, (2) stemming of the unnested tokens in order to reduce the morphological diversity of tokens (3) removal of grammatical stop-words, e.g. "I", "where", "that", retention of common n-grams as units within noun phrases, e.g., "climate change" becomes "climate_change", (4) following ngramisation, a list of commonly occurring personal stopwords are purged, e.g., "information" and "society" would be eliminated, but "information_society" would be retained if it proved to be a prevalent phrase.

Once the reasonably clean database is forged, it has to become a document-term matrix, where each document is a row, and each term a column, with the frequency of term-occurrence registered in every cell. The original DTM is far too sparse (i.e., contains too many low-frequency terms), and it also contains a few far too frequent terms. Sparsity reduction is required to eliminate low-frequency items—in our case, the original matrix has 1149 documents and 5142 individual terms and is more than 99% sparse, while after reduction, the DTM has dimensions of 1149 x 2024 and is 98% sparse. Finally, four omnipresent tokens were removed: "topic", "text", "public", and "topic_model". The LDA algorithm takes the DTM as input and requires the setting of several parameters, principally:

1. Hyperparameter delta—how likely it is for a token to belong to more than one topic, here we set delta at 0.01, which is moderately low.

2. Hyperparameter alpha—how likely it is for a document to be a mixture of more than one topic, here we set the initial alpha relatively low at 1.5, but we allow the model to estimate alpha further as it learns.

3. K-topic number: we tell our algorithm to find 14 topics.

RESULTS

An LDA model has two main outputs: (i) matrix beta and (ii) matrix gamma. The data analysis requires one or both of them, sometimes with metadata retained in the original abstract database. Since every document—an abstract in our case—is a mixture of topics, and every topic is a mixture of words—tokenised words after preprocessing, the beta matrix is created by extracting the per-topic-per-word probabilities for every topic/token combination. The tokens most strongly associated with each topic are featured in the faceted word clouds. Note that we already named the topics follow-ing extensive eyeballing of prominent tokens and publications associated with that topic. In naming topics, additional information was taken into account—other than the top-tokens, such as an examination of documents associated with the topic, the journals that most often publish articles associated with the particular topics, and the most prominent authors. Thus, the labels are not assigned by the model and constitute a hopefully well-informed judgement call on the part of the authors.

The other crucial model-output comes as the matrix "gamma", i.e., document—topic associations. The gamma value is essentially an estimate of the proportion of words from that document that belongs to the particular topic. Gamma is estimated for every document—topic pair; however, our alpha setting of the LDA hyperparameter pushed the algorithm towards looking for one dominant topic in every document. While multi-topicality is possible, even within the specific prose genre of article abstracts, we direct the model against pluralistic assignments for two reasons: (i) an abstract is a short-form document that is usually about one major topic, unlike, for instance, the full-text article which contains a mixture of topical threads, (ii) much of our further analysis relies on selecting the "top topic" of a document so for most purposes the multi-topical assignment of probabilities would be lost anyhow. In any case, this is a judgement call, and we calibrated the algorithm in this particular way after many attempts at modelling.

The fourteen topics identified and the main keywords are shown in Figure 1. At first glance, it is clear that the topics are diverse, but none of them include "sociology" or "sociological analysis" among the main keywords. In fact, the token "sociology" features only 41 times in the whole corpus. However, looking at the content of the topics, it is easy to identify issues that fall under the areas of specific sociologies (e.g. T2., T3. or T9.). A brief analysis of the documents most strongly associated with these three topics suggest, however, that they are typically conducted within other disciplinary frameworks. It seems indicative of the hitherto low level of engagement of sociological research with the novel methodologies of topic modelling. Further support is rendered to such a general observation by the examination of the most prominent journals publishing articles associated with the identified topics (see Table 1).

T1. Bibliometric analysis of academic journals T2. Public health outcomes climate_change covid learning journal knowledge discipline academic scholar scholar specific scientific scientific core central abstract source content reference database corporate government corporate government dirichlet allocation corporate government di map citation scholar government latent_dirichlet_allocation allocation corporate government stakeholder union discourse international energy national policy european focus citizen frame economic ^{law} environmental report understanding similarity frequency Ida antic Word latent task distribution T5. Attitudes to emerging technologies T6. Content classification and information extraction ies T6. Content classification and information extraction ent methodology tion unsupervised ant technique tec learning content diversity algorithm limitation originality algorithm attitude cross lin potential machine participation r nachine participation rele recommendation generation relevant T7. Social media analysis social_medium conversation medium twitter hashtag volume platform post sentiment disaster crisis relate tweet comment internet behavior mining green analytic experiment market online opinion detection communication T10. Spatial analysis and urban studies T9. News-media content analysis correlation crime region local distribution direct coverage candidate identity agenda correlation movement support power ideological election distribution dynamic temporal pattern ^{smart} urban city woman speech political we T11. Networks and network analysis T12. Identification of emerging trends and technologies account identification complex link dat_set dynamic collaboration similarity network collaboration complex link dida_set community community common distinct structure T14. Traffic analysis and urban commuting T13. Natural language processing technique qualitative mobility quantitative theme expert critical interpretation digital tool human corpus computational language rating tourism

Figure 1. Main topics within the topic modelling Source: own elaboration.

performance

customer

technique

The overall picture of existing uses of topic modelling suggests a strong diversity of interests. Some topics include specialised issues of data analysis using various machine learning algorithms methods (T4., T6., T13., as well as T1. and T6.), and others—detailed issues of new network technologies. Still, others are connected with customer services (T8.), transportation issues (T14.) or social media analysis (T7.).

Cluster analysis of the selected topics is presented in Figure 2. It clearly shows how some topics are connected to each other (e.g. T4. and T6.), while others are more separate entities (e.g. T9. or T7.). Note that the hierarchical clustering of topics is performed on data derived from the gamma-matrix, i.e., degree of association of each document with every topic. The most fundamental distinction occurs between the bottom five topics on the dendrogram, which broadly relate to methodological interests, and the other nine topics, which demonstrate more substantive concerns with specific knowledge domains.



Figure 2. Cluster analysis dendrogram of main topics. Source: own elaboration.

Apart from exploring the relationships between different topics, which can also be performed using more powerful classification techniques, it is also helpful to look inside the topics. Thus, a secondary LDA can be performed to identify micro-topics within the documents most strongly associated with each of the macro-topics. Figure 3 provides a snapshot of such an exploration. Notably, as implemented here, the analysis takes a number of analytic shortcuts—most importantly, all macro-topics are assumed to contain exactly eight micro-topics. It would be likely better to allow for macro-topics to have a variable number of micro-topics and make provisions for individualised hyper-parameter settings. Such an individualised approach would nevertheless require substantial manual data analysis within our current framework. Since secondary data users typically prefer analytic solutions which require minimal data-collection efforts on their part (Jabkowski, Cichocki, & Kołczyńska 2021), a onesize solution constitutes a preferable option to a manually tweaked one. The setting of secondary LDA parameters could be automatised in principle, but we have not yet achieved practically applicable solutions in this respect.

This Assertial in the specific processing in	ĩ	T14. Traffic analysis and - urban commuting	latent learning experiment resolution semantic	feature element recommendation service preference	topic_model real attraction exist occurrence	market online explanation movie post	incident technique base propose machine	transportation prevalence theme senior adas	usage performance source similarity algorithm	perception network event Ida category
17.2. Materialization intensional technologies Intelligence intensional intensio		T13. Natural language - processing	linguistic textual pattern methodology semantic	record resource chat academic librarian	public book metadata experiment reader	child reading art relate behavior	science interpretation insight document automate	critical archive journal algorithm content	word theme statistical genre school	quantitative scholar discourse thematic visualization
TI-1. Networks manuposities feet space		T12. Identification of emerging - trends and technologies	intelligence direction knowledge artificial structure	cycle topic_model planning life dynamic	past latent detail base journal	mining word path main document	business financial company disclosure hot	competency skill education popular success	network key ent_dirichlet_allocatie empirical performance	on growth methodology uavs
HD0 Source urbs studies Source wader urbs studies Source urbs studies		T11. Networks and network - analysis	text knowledge radical specific algorithm	latent international link detection data_set	authorship science climate interdisciplinary theoretical	online labor communication comment identification	property cooperation node hierarchical bayesian	public interaction connection medium synthetic	tweet feature opinion technique cluster	diversity computational transition potential experiment
T9. News- media content- section of the condition personal massing Letter massing Petter media content- section of the condition personal massing Petter massing Petter media personal section of the condition personal section of the condition Petter massing Petter personal massing Petter personal massing<	corpus	T10. Spatial analysis and - urban studies	source vulnerability word vector document	risk quality water building tourism	similarity propose trip mining technique	geo event dynamic temporal public	interaction topic_model obesity theme detection	land classification mobility feature river	medium transportation community specific government	learning methodology correlation limitation cuber
TS. Customer and market and market	n the entire	T9. News- media content - analysis	text ideological topic_model specific speaker	election candidate personal russian post	sexual woman popular harassment language	platform government event toxic page	newspaper outlet structure partisan report	politician legislative immigration populist electoral	public discourse network online hate	comment audience sentiment identity religious
T7. Social media analysis codd georenial media matysis platform pattern pattern media matysis platform pattern pattern pattern media matysis platform pattern pattern pattern pattern platform pattern metwork share platform pattern metwork platform pattern platform paperinent platform pattern <	performed or	T8. Customer experience _ and market analytics	natural certify latent accommodation development	performance company provider methodology behavior	employee current infrastructure policy unstructured	content restaurant trip advisor attribute	insight business technique tweet retailing	public city comment medium base	website news online_review opportunity recommendation	mining park opinion visitor theme
T6. Content classification and information extraction Electronic propose generation Categoory probability poundities Collection istinating textual news istinating generation teg relevant propose_method orgoph without generation service istinating textual T5. Attitudes to emerging- technologies comparative instructure cross resource comparative instructure cross cross optition propose_method potitical exist platform exist platform exist platform to this propose_method istructure interaction comparative event compute propose_method document interaction interaction comparative cross T4. Machine- mining conserved interaction potitical events to thick platform to thick probabilition generation for propose_method document interaction document comparative cross document interaction document propose_method document interaction interaction interaction T4. Machine- mining coherence indection modulingual interaction short conserved interaction potical event probabilition generation potical event probabilition generation probabilition generation protoce interactionic generation for propose interactionic generation interactionic event comparative interactionic generation interactionic event comparative interactionic generation interactionic event comparative interactioni generation T3. Public interact	10): LDA F	T7. Social _ media analysis	covid discourse potential misinformation pandemic	platform pattern real source geotagge	network political narrative topic_model share	stage public_opinion management Ida text	topic_model engagement damage library awareness	word online micro blog emerge	communication natural refugee relief demand	emotion detection health story semantic
Upper T5. Attitudes to emerging - technologies comparative minimaging - real patterm originality sainetwork diversity political platform development text ida authorship vool structure whole anxiety tool semantic whole anxiety krowledge procedure semantic generation preference document community polential machine iberaction composite game polential machine 14. Machine- aided text - mining coherence allocation supervised for sainetwork diversity short translation for courrence pattern selection for whist polential authorship tool matrix human structure generation proor matrix human structure collection weetor weet polential machine sense knowledge weepriment allocation proor sense weepriment allocation 13. Public policy analysis policy analysis acatement selection for sphere relate environmental agaptation content agenda climate sphere erigein focus tooling university newspaper structure relate environmental ports development topic model report generation international news specific international gootinn cerce ontimunication content cerce cite chie sphere erigein focus tooling university newspaper international text topic, model gootal adoptation international gootal gootal adoptation for specific for specific for specific for specific for specific for specific 1 1 2 3 4 5	o-topics (k =	T6. Content classification and information extraction	learning wikipedia structure core generation	category network relate probability journal	distribution multiple latent training textual	collection similarity Ida short segment	news topic_model experiment page engine	tag relevant rank respect resource	graph extraction vector summarization propose_method	service knowledge technique event corpus
T4. Machine- aided text - mining coherence allocation multilingual language allocation short courrence patent selection point retrieval categorization matrix human structure collection vector vec network algorithm sense experiment algorithm T3. Public policy analysis fstakeholder content agenda climate sphere relate content agenda climate sphere relate environmental agenda structure relate environmental agenda development topic ross network agency bealth international news specific content international agency topic international region cosi international topic ross international topic ross international region cosi international topic ross international topic ros	Macro	T5. Attitudes to emerging - technologies	comparative misinformation language cross methodology	pattern originality sai network diversity	political exist platform medium development	text Ida authorship validity tool	structure vehicle anxiety knowledge procedure	semantic generation technique latent preference	document cqa answer potential machine	interaction behavior complete learning real
T3. Public polcy analysis - polcy analysis - polcy analysis - gold content agenda with content of sphere frequence of the environmental newspaper structure region focus crisis development big crisis international content operation of the environmental newspaper international content operation of the environmental newspaper structure region focus crisis development big crisis international content operation of the environmental newspaper international content operation of the environmental newspaper science for the environmental content operation of the environmental newspaper science for the environmental operation of the environmental operation o		T4. Machine- aided text - mining	coherence accuracy improvement baseline supervised	multilingual language allocation translation cross	short content occurrence pattern selection	exist probabilistic generation embedding inference	patent keyword real retrieval categorization	matrix human structure collection prior	vector vec network algorithm embed	sense knowledge experiment bag ent_dirichlet_allocation
T2. Public health outcomes medium mental adaptation dacptation dacptation discourse theme life upper umental adaptation current text topic pornal text support per corona word tweet sentiment air machine post forum interaction sentiment content drug nurse government news college teaching teaching sclour T1. Bibliometric journals development interaction pattern heme scholar structure cite dynamic word specific philosophy abstract relate topic model biometric computational document title research, topic map source sociology special metwork research, topic university methodology methodology methodology methodology methodology methodology abstract i i i i i j i j i j		T3. Public _ policy analysis	stakeholder political content agenda variation	climate sphere relate environmental newspaper	structure region focus funding crisis	development topic_model report agency health	national news understanding specific collection	international law text network opinion	economic communication coal university rural	science citizen sustainability technique natural
T1. Bibliometric analysis- of academic journalsdevelopment international education pattern themescholar structure cite dynamicspecific evolution philosophy abstract relatetopic model bibliometric computational core miningtitle network tesact h. topic human core miningsociology metwork thematic is map sourceuniversity research h. topic human communitysociology metwork thematic is map sourceuniversity research human communityuniversity metwork communityiii		T2. Public _ health outcomes	medium mental global adaptation discourse	theme life language woman current	text topic_model community school journal	disease support peer corona word	tweet quality sentiment air machine	post forum international content	drug nurse government news college	teaching teacher science development key
$\frac{1}{1}$ $\frac{1}{2}$ $\frac{1}{3}$ $\frac{1}{4}$ $\frac{1}{5}$ $\frac{1}{6}$ $\frac{1}{7}$ $\frac{1}{8}$ Micro-tonics (k = 8): 1 DA performed on subsets of documents associated with each magna tonic		T1. Bibliometric analysis - of academic journals	development international education pattern theme	scholar structure cite dynamic word	specific evolution philosophy abstract relate	topic_model bibliometric computational core mining	document thematic lis map source	title network research_topic human community	sociology special sport central similarity	university methodology main communication digital
$\mathbf{x}_{1} = \mathbf{x}_{1} + \mathbf{x}_{2} + \mathbf{x}_{1} + \mathbf{x}_{2} + \mathbf{x}_{2} + \mathbf{x}_{3} + \mathbf{x}_{1} + \mathbf{x}_{2} + \mathbf{x}_{1} $			i	2 Micro terrico	$\frac{1}{3}$	4 formed on subsets	5	6 opiated with each		8

Topics-within-topics: micro-topic modelling within each of the macro-top	oics
--	------

Figure 3. Secondary LDA: micro-topics within macro-topics. Source: own elaboration.

Торіс	Journal			
T1. Bibliometric analysis of academic journals	Journal of the Association for Information Science and Technology, Journal of Con- sumer Research, Cognition, International Journal of Communication, Environmental Sociology			
T2. Public health outcomes	Global Environmental Change, Journal for Research in Mathematics Education, European Societies, Scientometrics, Journals of Gerontology –Series B Psychological Sciences and Social Sciences			
T3. Public policy analysis	Policy and Society, Sustainability, Communication Methods and Measures, Policy Studies Journal, Journal for the Scientific Study of Religion			
T4. Machine-aided text mining	Information Retrieval, Scientometrics, Topics in Cognitive Science, Communication Methods and Measures, Journal of Information Science			
T5. Attitudes to emerging technologies	Political Analysis, Information Processing and management, Social Science Comput- er Review, Transportation Research Part A: Policy and Practice, Decision Support Systems			
T6. Content classification and information extraction	Information Processing and Management, Decision Support Systems, Scientomet- rics, Synthesis Lectures on Human Language Technologies, Information Retrieval			
T7. Social media analysis	Decision Support Systems, Computers, Environment and Urban Systems, Interna- tional Journal of Information Management, Journal of Information Science, Digital Journalism			
T8. Customer experience and market analytics	International Journal of Information Management, Tourism Management, Journal of Service Research, Decision Support Systems, Social Science Computer Review			
T9. News-media content analysis	American Journal of Political Science, Political Analysis, Information Communi- cation and Society, Political Behavior, European Journal of Cultural and Political Sociology			
T10. Spatial analysis and urban studies	International Journal of Geographical Information Science, Computers, Environment and Urban Systems, Cartography and Geographic Information Science, Transporta- tion Research Part C: Emerging Technologies			
T11. Networks and network analysis	Journal of Informetrics, Decision Support Systems, Scientometrics, Journal of the Association for Information Science and Technology, Resources Policy, Review of International Organizations, Social Science Computer Review			
T12. Identification of emerging trends and tech- nologies	Information Processing and Management, International Journal of Engineering Edu- cation, Migration Studies, Futures, Telecommunications Policy			
T13. Natural language pro- cessing	Digital Journalism, Poetics, Digital Scholarship in the Humanities, Administrative Science Quarterly, Government Information Quarterly, International Journal on Digital Libraries			
T14. Traffic analysis and urban commuting	Transportation Research Part C: Emerging Technologies, Accident Analysis and Prevention, Computational Linguistics, Tourism Management, Journal of Air Trans- port Management			

Table 1. Topics and key journals Source: own elaboration.

Modelling micro-topics within the fourteen macro-topics serves two main goals: (1) investigation of the macro-topic coherence, (2) exploration of the within-topic diversity of interests. The former use is methodological and may be deployed to evaluate the quality of macro-topics. Thus, it complements other measures of model fit, e.g., topic coherence or model perplexity, as well as those of manual inspection, e.g., a review of top documents or top tokens. In this methodological use, the tool is most restricted by the above mentioned fixed topic number and parameter settings for the micro-topic modelling. However, it seems to work best as a snapshot of discourse, allowing an inspection of topic diversity at a glance. Such an approach best fits exploratory studies, which aim to gain quick insights into an unknown body of text. For instance, the algorithms deployed here to study abstracts mentioning "topic modelling" could easily be re-purposed to analyse abstracts relating to any other research domain. Even a cursory examination of the micro-topics points to the existence of meaningful micro-topics. For instance, within T.7. Social media analysis there are several recognisable themes: T.7.1. "Covid pandemic (mis)information", T.7.3. "Network dissemination of political narratives" or T.7.7. "Communication patterns of refugees".

As mentioned above, the LDA algorithm does not have access to any of the publication meta-data—modelling only involves document ids and abstract texts. However, once the documents are classified as belonging to any particular topic, this information can be merged with the available meta-data. For example, it is possible to determine which journals provide the most prominent publications within each topic. Detailed information on the journals assigned to the particular topics can be found in Table 1. As can be seen, topics have a heterogeneous representation of journals, which means that topic modelling itself is used in different journals assigned to specific disciplines. Looking from another perspective, although there is no "sociology" in the keywords, as we pointed out above, there are sociology (and multidisciplinary with sociology) journals in which the machine learning algorithms of topic modelling are applied (cf. *Environmental Sociology, European Societies, European Journal of Cultural and Political Sociology, Migration Studies*).

CHALLENGES AND LIMITATIONS

When considering the role of topic modelling within the development of sociology, which to explain social reality increasingly conditioned by digital technologies must develop adequate methods of analysis, one cannot ignore the diversity of methods of analysing large corpora of data as their weaknesses. This paper is based on an implementation of LDA (Blei, Ng, & Jordan 2003), but there are also other established alternatives (Bohr & Dunlap 2018). Table 2 provides a brief discussion of four topic modelling methods along with their limitations.

Name of the methods	Characteristics	Limitations	
Latent Semantic Analysis (LSA)	LSA can get from the topic if there are any synonym words.	It is hard to obtain and to de- termine the number of topics.	
	Not robust statistical background	To interpret loading values with probability meaning, it is hard to operate it.	
Probabilistic Latent Se- mantic Analysis (PLSA)	It can generate each word from a sin- gle topic; even though various words in one document may be generated from different topics.	At the level of documents, PLSA cannot do probabilistic model.	
	PLSA nanules polysenty.		
Latent Dirichelet Alloca- tion (LDA)	Need to manually remove stop-words. It is found that the LDA cannot make the representation of relationships among topics.	It becomes unable to model re- lations among topics that can be solved in CTM method.	
Correlated Topic Model (CTM)	Using of logistic normal distribution to create relations among topics. Allows the occurrences of words in other topics and topic graphs.	Requires lots of calculation. Having lots of general words inside the topics.	

Table 2. The characteristics and limitations of topic modeling methods Source: Alghamdi & Alfalqi 2015: 150–151.

The LDA example shows that the listed characteristics and limitations are neither complete nor definitive. For example, the need to manually remove stop-words remains a problem even after advanced pre-processing. The high-frequency stop-words are not a problem, i.e., for most common languages, there are well-researched dictionaries available. Domain-specific stop-words may prove challenging. For instance, when dealing with academic abstracts, it seems reasonable to exclude such common words as: "paper", "study", "article", "issue", "research", "analysis", "finding", "approach", "author", "program", "review' or "chapter". Such words constitute a common occurrence in abstracts, regardless of their topic. On the other hand, there are common methodological words as "logistic", "regression", "multilevel", "hypothesis", or "regression model"; they also are omnipresent in academic abstracts regardless of their research interests, it could be argued that this particular set of tokens would indicate that the underlying research is quantitative. Hence, the precise set of stopwords remains of the author's making and usually involves multiple trial-and-error iterations. For more information on the limitations of topic models, those interested can find quite a lot of literature on the subject (Arabshahi & Anandkumar 2016; Blei & Lafferty 2006; Ding & Jin 2019; Lee, Song, & Kim. 2010).

CONCLUSIONS

The use of value-laden categories (cf. Gans 1999: 268; Rex 1983; Weber 1949) to evaluate sociology has, on the one hand, a heuristic dimension aimed at drawing attention to the problem of the quality of sociological research rather than the normative evaluation of the discipline as a whole. However, on the other hand, a "good" sociology primarily describes and explains, and to a lesser extent predicts, the functioning and changes of society. A "good" sociology allows us to see the invisible or to explain the known in a different, often non-intuitive way. "Bad" sociology, on the other hand, is not sociology based on coding (as stated by Lee & Martin 2015), but an approach that is unable to diagnose and explain social reality. In this view, topic modelling based on LDA, although it has limitations of (i) applicability and of (ii) a methodological nature, enriches the sociological approach by enabling the analysis of large textual datasets that would not be possible without this method (DiMaggio, Nag, & Blei 2013: 577). It constitutes a relatively easy to use method for investigating textual Big Data, which remains difficult or impossible to grasp through traditional empirical approaches.

Additionally, "good" sociology using topic modelling can serve as, on the one hand, a novel method of literature review, initiating further research using "classical" research methods. Principally, it can serve as a discourse-mapping tool, identifying areas of interest and potential coding schemes for more conventional analysis. As it can be deployed rapidly at scale, it seems to constitute a good fit for meta-analyses of literature and exploratory summarisation of prevailing trends. One such ready opportunity exists in the form of secondary LDA (LDA-within-LDA-results), which has been demonstrated in this paper. On the other hand, LDA can be deployed as a fully holistic research method both on its own and in conjunction with other meta-data (e.g., monitoring topic prevalence over time, tracking funding sources for particular research streams, or investigating publication patterns). Such research opportunities were demonstrated in this paper, as we identified which journals publish most prominently within each of the identified macro-topics.

FUNDING: This work was supported by grants awarded by the National Science Centre, Poland [no. 2021/05/X/HS6/00067] and [no. 2018/31/B/HS6/00403].

CONFLICT OF INTEREST: The authors declare no conflict of interest.

REFERENCES

- Adorjan, Michael & Benjamin Kelly. 2021. "Time as Vernacular Resource: Temporality and Credibility in Social Problems Claims-Making." *The American Sociologist* 1–27. <u>https://doi.org/10.1007/s12108-021-09516-x</u>
- Alghamdi, Rubayyi & Khalid Alfalqi. 2015. "A survey of topic modeling in text mining." *International Journal of Advanced Computer Science and Applications* 6(1): 147–153.
- Arabshahi, Forough & Animashree Anandkumar. 2016. Beyond LDA: A unified framework for learning latent normalized infinitely divisible topic models through spectral methods. Technical report. Retrieved November 10, 2021 (https://escholarship. org/content/qt7d95h1dd/qt7d95h1dd_noSplash_f43b4c2f867fcf6945df3700d-0196f3a.pdf).

- Baranowski, Mariusz. 2021. "The sharing economy: Social welfare in a technologically networked economy." *Bulletin of Science, Technology & Society* 41(1): 20–30. https://doi.org/10.1177/02704676211010723
- Baranowski, Mariusz & Dorota Mroczkowska. 2021. "Algorithmic Automation of Leisure from a Sustainable Development Perspective." Pp. 21–38 in Handbook of Sustainable Development and Leisure Services. World Sustainability Series, edited by A. Lubowiecki-Vikuk, B. M. B. de Sousa, B. M. Đerčan, & W. Leal Filho. Cham: Springer. https://doi.org/10.1007/978-3-030-59820-4_2
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, & Akitaka Matsuo. 2018. "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software* 3(30): 774. doi:10.21105/ joss.00774
- Berelson, Bernard R. 1952. *Content analysis in communication research*. Glencoe, Ill.: Free Press.
- Blei, David M., Andrew Y. Ng, & Michael I. Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3(1): 993–1022.
- Blei, David & John Lafferty. 2006. "Correlated topic models." *Advances in Neural Information Processing Systems* 18: 147.
- Bohr, Jeremiah & Riley E. Dunlap. 2018. "Key Topics in environmental sociology, 1990–2014: Results from a computational text analysis." *Environmental Sociology* 4(2): 181–195. DOI: 10.1080/23251042.2017.1393863
- DiMaggio, Paul, Manish Nag, & David Blei. 2013. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding." *Poetics* 41(6): 570–606. <u>https://doi.org/10.1016/j.poetic.2013.08.004</u>
- Ding, Juncheng & Wei Jin. 2019. "A Prior Setting that Improves LDA in both Document Representation and Topic Extraction." 2019 International Joint Conference on Neural Networks (IJCNN) 2019: 1–8. DOI: 10.1109/IJCNN.2019.8852000
- Gans, Herbert J. 1999. *Making Sense of America: Sociological Analyses and Esseys*. Lanham, Oxford: Rowman & Littlefield Publishers, Inc.
- Gouldner, Alvin W. 1976. *The Dialectic of Ideology and Technology: The Origins, Grammar, and Future of Ideology*. New York: Seabury Press.
- Grün Bettina & Kurt Hornik. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40(13): 1–30. doi: 10.18637/jss.v040.i13
- Hannigan, Timothy R. et al. 2019. "Topic modeling in management research: Rendering new theory from textual data." *Academy of Management Annals* 13(2): 586– 632.
- Jabkowski, Piotr, Piotr Cichocki, & Marta Kołczyńska. 2021. "Multi-Project Assessments of Sample Quality in Cross-National Surveys: The Role of Weights in Applying External and Internal Measures of Sample Bias." *Journal of Survey Statistics and Methodology* 1–24. <u>https://doi.org/10.1093/jssam/smab027</u>
- Lasswell, Harold D. 1927. "The theory of political propaganda." *American Political Science Review* 21(3): 627–631.
- Lazarsfeld, Paul F. & Anthony R. Oberschall. 1965. "Max Weber and Empirical So-

cial Research". *American Sociological Review* 30(2): 185–199. <u>https://doi.org/10.2307/2091563</u>

- Lee, Sangno, Jaeki Song, & Yongjin Kim. 2010. "An empirical comparison of four text mining methods." *Journal of Computer Information Systems* 51(1): 1–10. DOI: 10.1080/08874417.2010.11645444
- Lee, Monica & John L. Martin. 2015. "Coding, counting and cultural cartography." *American Journal of Cultural Sociology* 3(1): 1–33. https://doi.org/10.1057/ajcs.2014.13
- Mayntz, Renate, Kurt Holm, & Peter Hübner. 1976. *Introduction to Empirical Sociology*. Harmondsworth: Penguin Education.
- McFarland, Daniel A., Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D. Manning, & Daniel Jurafsky. 2013. "Differentiating language usage through topic models." *Poetics* 41(6): 607–625. <u>https://doi.org/10.1016/j.poetic.2013.06.004</u>
- Mohr, John W. & Petko Bogdanov. 2013. "Introduction—Topic models: What they are and why they matter." *Poetics* 41(6): 545–569. <u>https://doi.org/10.1016/j.poetic.2013.10.001</u>
- Pääkkönen, Juho & Petri Ylikoski. 2021. "Humanistic interpretation and machine learning." *Synthese* 199: 1461–1497. https://doi.org/10.1007/s11229-020-02806-w
- R Core Team. 2021. *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria. <u>https://www.R-project.org/</u>
- Rex, John. 1983. "British Sociology: 1960-80—An Essay." *Social Forces* 61(4): 999-1009. https://doi.org/10.2307/2578275
- Schmiedel, Theresa, Oliver Müller, & Jan vom Brocke. 2018. "Topic modeling as a strategy of inquiry in organizational research." *Organizational Research Methods* 22(4): 941–968. <u>https://doi.org/10.1177/1094428118773858</u>
- Selwyn, Neil. 2015. "Data entry: Towards the critical study of digital data and education." *Learning, Media and Technology* 40(1): 64–82. https://doi.org/10.1080/174 39884.2014.921628
- Silge, Julia & David Robinson. 2017. *Text mining with R: A tidy approach*. Sebastopol, CA: O'Reilly Media.
- Weber, Max. 1949. "'Objectivity' in Social Science and Social Policy." Pp. 50–112 in *Max Weber on The Methodology of the Social Sciences*, edited by E. A. Shils & H. A. Finch. Illinois: The Free Press of Glencoe.

Weber, Robert Philip. 1990. Basic content analysis. London: Sage.

Wilterdink, Nico. 2012. "Controversial science: Good and bad sociology." *Figurations: Newsletter of the Norbert Elias Foundation* 36: 1–12. <u>https://pure.uva.nl/ws/</u> <u>files/4493017/151330_380325.pdf</u>

BIOGRAPHICAL NOTE

Mariusz Baranowski is assistant professor of sociology at the Adam Mickiewicz University, Poznań, Poland.

Piotr Cichocki is assistant professor of sociology at the Adam Mickiewicz University, Poznań, Poland.

OPEN ACCESS: This article is distributed under the terms of the Creative Commons Attribution Non-commercial License (CC BY-NC 4.0) which permits any non-commercial use, and reproduction in any medium, provided the original author(s) and source are credited.

ARTICLE HISTORY: Received 2021-12-01 / Accepted 2021-12-30