Journal of Informetrics 16 (2022) 101224

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Robustness, replicability and scalability in topic modelling

Omar Ballester*, Orion Penner

École Polytechnique Fédérale de Lausanne (EPFL) Lausanne, Switzerland

ARTICLE INFO

Keywords:
Scientometrics
Topic modelling
Stability
Robustness
Similarity
Informetrics

ABSTRACT

Approaches for estimating the similarity between individual publications are an area of long-standing interest in the scientometrics and informetrics communities. Traditional techniques have generally relied on references and other metadata, while text mining approaches based on title and abstract text have appeared more frequently in recent years. In principle, topic models have great potential in this domain. But, in practice, they are often difficult to employ successfully, and are notoriously inconsistent as latent space dimension grows. In this manuscript we identify the three properties all usable topic models should have: robustness, descriptive power and reflection of reality. We develop a novel method for evaluating the robustness of topic models and suggest a metric to assess and benchmark descriptive power as number of topics scale. Employing that procedure, we find that the neural-network-based paragraph embedding approach seems capable of providing statistically robust estimates of the document–document similarities, even for topic spaces far larger than what is usually considered prudent for the most common topic model approaches.

1. Introduction

Methods for understanding individual documents' topics and concepts—such as patents or scientific publications—are a matter of long-standing interest within the scientometric and informetric communities. Indeed, going back to some of Garfield's earliest thinking on citation indexes (1955), he identified a goal of an "association-of-ideas" index. In those thoughts, he further developed the role such an index would play in the literature-search process. He highlighted the value of a "sub-micro" or "molecular" level approach over one focused on "classification".

Today, document similarity and clustering is a vibrant area of research within the scientometric community, expanding to many areas of application within the social sciences. Examples include information retrieval, the mapping of science, and metrics to enrich studies of the individuals and institutions engaged in the research production process. Much of today's work, in line with Garfield's early vision, place citations and co-citations at the centre of their formulation of contextual similarity. However, the relationship between citations and topics may be more tenuous than generally accepted (Borner et al., 2003).

Increases in computational capacity and the availability of electronic data have opened many new avenues for estimating document similarity and have enabled clustering of documents. Pioneers in the field began to use hybrid co-word and co-citation analysis as early as the 1990s (Braam et al., 1991; Callon et al., 1991; Glenisson et al., 2005; Noyons and van Raan, 1994). While the range of options and ideas is vast, in this manuscript we focus on "Topic Models"—a group of techniques arising mainly from the computer science literature, but already widely used in science characterisation (Boyack et al., 2011; Glaser et al., 2017). As the input to these techniques is textual data (a collection of text documents), they offer a twist on traditional approaches for understanding the topics

* Corresponding author at: École Polytechnique Fédérale de Lausanne (EPFL), Sécrétariat Géneral BS 250 - Station 4 CH-1015 Lausanne Switzerland.

E-mail addresses: omar.ballester@epfl.ch (O. Ballester), orion.penner@epfl.ch (O. Penner).

https://doi.org/10.1016/j.joi.2021.101224
Received 11 May 2021; Received in revised form 17 October 2021; Accepted 4 November 2021
Available online 20 November 2021

and concepts that make up individual publications and, in turn, for estimating document similarities and clustering. These techniques are certainly not without their flaws (Velden et al., 2017), but they are also well positioned to exploit the rapidly growing body of textual data and perhaps even full text.

In recent years, stochastic topic models have seen widespread adoption and multiplicity of applications. These models, however, suffer from systemic errors due to topic instability driven by the initialisation stage (Belford et al., 2017), slight variations in data (Hecking and Leydesdorff, 2018) or just document-reordering upon training (Agrawal et al., 2018). Our contribution goes beyond prior art in testing stochastic topic models in that we study the stability across runs. Besides the stability, there is congruity amongst scholars from different knowledge areas that there is a trade-off between interpretability and predictive power (Blei et al., 2010). As a result, as topics become increasingly fine-grained, they improve their predictive likelihood but become less useful for human interpretation (Chang et al., 2009).

For these two problems, we propose an evaluation that is not data-dependent nor requires “human” interpretation. As well, our assessment of stability and scale can extend to neural-network-based models such as Word2Vec/Doc2Vec. The contribution is twofold. First, we provide an answer to the problem of self-evaluation of topic models for research applications in the absence of gold-standards. Second, we provide a metric for the stability of subsequent runs of topic models—in other words, we measure the dispersion of arbitrariness created by idiosyncratic solutions of topic models.

In this manuscript, we identify the three properties that functional topic models should have: robustness, descriptive power and reflection of reality. We test and compare the performance of NMF, LDA and Doc2Vec regarding their scalability to high dimensions (the descriptive power they hold) and their consistency (robustness) across estimations. We corroborate that the models we train reflect reality, but leave further discussion on realism out of the scope of this paper. We carry out this analysis on scientific bibliographic data, but the concepts, methods, and implementation are easily extended to any social-science research involving text data.

The remainder of the paper is structured as follows. In Section 2, we introduce the most common topic modelling techniques and their applications to information science, and we describe the data of our analysis. In Section 3 we describe our method of evaluation of topic models, including the training set-up and the metrics of interest. Section 4 discusses the results, and their limitations. Finally, in Section 5 we discuss potential applications of neural-network-based topic models and provide examples of successful applications.

2. Methods and data

2.1. Topic models

Topic models are statistical models designed to extract from a set of documents the relevant “topics”. In turn, topic models represent each document within that “topic” or latent space. The most common application of topic models to the characterisation of science and innovation use probabilistic topic models (Boyack et al., 2011; Glaser et al., 2017). Griffiths and Steyvers (2004) were the first to apply Latent Dirichlet Allocation (LDA) to article abstracts in order to find scientific topics and illustrate the contextual relationships between different disciplines. Later researchers have applied probabilistic models to many different levels of aggregation (Hall et al., 2008; Lu and Wolfram, 2012; Rosen-Zvi et al., 2010; Tang et al., 2008). However, Blei et al. (2010) found that, often, practitioners directly assume that the latent spaces (topic space) generated by the model are semantically meaningful without an in-depth quantitative evaluation (Chang et al., 2009; Hall et al., 2008; Mei et al., 2007; Newman et al., 2010).

Specifically, we conduct a comparison of three topic models: Non-negative Matrix Factorisation (NMF) (Lee and Seung, 1999), LDA (Blei et al., 2003)— two widely used approaches in the literature—and paragraph embeddings (Le and Mikolov, 2014). NMF decomposes the document-term matrix into a product of two matrices that contain only non-negative entries by definition. LDA is based on a probabilistic model of language. Through matrix decomposition, LDA produces two matrices stochastically. Paragraph embeddings is a relatively novel neural-network-based approach built upon the similarly new Word2Vec word-embedding algorithm by Mikolov et al. (2013). Word2Vec formulates the problem as one of predicting an omitted word within a short (3 to 15 words) contiguous sequence of text.[1] Treating the neural network's hidden layer as the latent space, one can infer document-topic couplings from the model's parameters. Although, strictly speaking, Doc2Vec may not be considered a “true” topic model as the topic-term couplings inference is not straightforward. However, this feature (or lack thereof) is acceptable as the fundamental elements of the statistical analysis presented herein are document-document similarity scores, which require only the document-topic vectors.

Most, if not all, topic models require the user to define the size of the “topic” (or latent) space. The question of what is the “best” or optimal size of the topic space comes up often in the literature, and no clear criteria exist (Glaser et al., 2017). In theory, increasing the size of the latent space increases the granularity in which the model represents ideas. Despite the many exciting lines of research that could potentially be tackled by pushing topic models to high dimensional topic spaces, topic models are rarely employed with a latent dimension greater than, perhaps, a few dozen. Often, if the granularity is pushed too high, the topics start to degrade into incoherent nonsense (Greene et al., 2014; Hecking and Leydesdorff, 2018; Steyvers and Griffiths, 2013). In plain language, at some point, the exact same algorithm, with the exact same parameters, on exactly the same data, but with a different random seed, will produce a quantitatively and qualitatively different set of document-topic and topic-term vectors (Belford et al., 2017).

The topic modelling literature proposes information-theoretic measures to estimate the extent to which topic-term vectors vary from run to run. However, it is indeed the case that changes in the topic-term vectors may *not* preclude stability when considering

[1] Thus, this is not, strictly speaking, a bag-of-words approach while NMF and LDA are.

only document-document similarities. That is, even if the topics themselves are inconsistent from one training to the next, the measure of pairwise similarity may not change.

In our analysis, we use a Python implementation of NMF and LDA from the sci-kit learn library (Pedregosa et al., 2011). For paragraph embeddings, we use Python's Doc2Vec (Řehuřek and Sojka, 2010), which follows the architecture laid out by Le and Mikolov (2014). In particular, we use the Distributed Bag of Words (DBOW) approach with negative sampling, as elaborated in Levy et al. (2015) and Dai et al. (2015).

2.2. Data

In topic models, one transforms document-term matrices into document vectors that exist in the latent-space. For our training, a *document* is the career output of a researcher, and the *terms* are Medical Subject Headings (MeSH). For each researcher, we extract all assigned MeSH from their publications. We rely on the 2014 version of PubMed, which provides the individual publication metadata, from which we use Journal, Year, Document Type and Medical Subject Headings. We identify the output of each researcher from the Author-ity disambiguation of PubMed carried out by Torvik and Smalheiser (2009).[2]

The document-term vector resulting from this procedure is one in which each vector entry corresponds to the concatenated list of MeSH terms assigned to the given researcher's publications across the entirety of her career. Disambiguation is thus crucial for the reliability of the models. The Author-ity disambiguation achieves a recall of 98.8 percent, with fewer than 2 percent of the author's production split into two and fewer than 0.5 percent of distinct authors clustered as one. We deal with careers starting 1974 or later, noting that our data terminate in 2009 as that is the maximum extent of the disambiguation data. In order to work with comparable, large enough documents, we filter-out researchers with fewer than 50 research publications and fewer than ten years of activity, which yields about 147,000 researchers (documents) for analysis. As well, we exclude all publications that are not classified as research articles. Each "document" (researcher) contains, on average, 930 non-unique tokens that come from an average of 118 distinct publications. Our corpus of publications contains MeSH from over 17 million scholarly articles. In turn, each publication contributes about eight MeSH terms randomly sorted (to break the alphabetic relationship) upon introduction in the term list. There are 22,909 unique MeSH terms.

MeSH have several advantages for applying topic modelling: (i) They are standardised (both in spelling and scientific terms—*i.e.* use of the prevalent terms for neoplasms, cell nomenclature, or technique names. (ii) MesH terms are harmonised. They suppress the common problems of free text, including synonyms, term permutations and acronyms all referring to the same subject. (iii) Unlike author-keywords, they are assigned by a centralised agency, reducing the self-selection bias. (iv) They are content descriptors. (v) MeSH terms include multi-word tokens, increasing the specificity while providing a different token for distinct or narrower concepts—e.g.: "Cell" or "Stem Cell" or "Embryonic Stem Cell". (vi) They deal with term diachronicity.

Therefore, MeSH terms represent a meticulously curated description produced by a third party (alien to the authors of the manuscript or journal). The result is a unified set of keywords that encapsulate the article's content they epitomise with enough precision to classify the knowledge that the publication covers. While the use of MeSH terms does not fully leverage the power of the most recent natural language processing (NLP) techniques, it provides a controlled and curated vocabulary, which largely simplifies the pre-processing stage. In this paper, we compare the performance of different topic modelling techniques, so the more restrictive vocabulary set plays to our advantage. First, it reduces the document-term matrix dimension, and second, the technical stages are shortened by eliminating the burden of text tokenisation. The data is constantly updated, as terms are outdated and new synonyms used. MeSH term introduction is conservative, and does not happen until the community has adopted a standard term.

We include a heuristic rule of topic classification based on the journal of the articles. For this, we use Eigenfactor's journal classification labels (Bergstrom et al., 2008; Rosvall and Bergstrom, 2008), identifying the subject family of all our indexed journals. We do not use the rule-based classification in this manuscript for any purpose other than sub-sampling the database or establishing comparison groups. We do not use the labels as ground truth for topic generation or incorporate them in any classification pipeline. See the Supplementary Material, Section 1 for more details.

3. Analysis

In the scientific literature, topic models are seldom used as an all-in-one, out-of-the-box topic identification tool. Rather, topic models act as an intermediate step to facilitate large text corpora's readability and interpretability by humans. Reflecting upon the gaps highlighted by many scholars before us, we have identified three key properties that topic models must demonstrate. To be suitable for application in the social sciences, models must exhibit:

1. Statistical robustness. Running the same model on the same data with the same parameters should produce the same results—or at least highly similar.
2. Descriptive power that increases with the size of the latent dimension. Changing the number of topics should alter the results both qualitatively and quantitatively. Document granularity should follow the changes in topic space size.

[2] We acknowledge this choice of document is not conceptually the best representation of a researcher's career—multiple documents that account for different career stages would represent a career better. However, the focus of the paper lies on the reproducibility of topic modelling, rather than a particular solution to modelling research careers, and this data organisation serves this purpose best.

3. Reflect reality. The results produced by topic modelling, document-document similarities or clustering or otherwise, must be consistent with patterns and relations known to exist within and across research domains.

Below we propose and execute specific statistical tests concerning the first two, while for the third, we provide preliminary evidence and highlight paths for further work.

3.1. Statistical robustness

Given the non-linear nature of probabilistic topic models, there is no *a priori* ordering that makes the topics identifiable between runs of the algorithm (Steyvers and Griffiths, 2013). That means that the space-vector resulting of two subsequent runs is not necessarily the same, and the document representation is, therefore, difficult to compare. It is then necessary to know which topics are stable and not idiosyncratic to a particular solution.

The majority of efforts to evaluate coherence in probabilistic topic models have concentrated around *expert evaluation* (Blei et al., 2010) or *entropy* metrics — *i.e.* information theory metrics closely linked to entropy, such as Pointwise Mutual Information (Velden et al., 2017), symmetrised Kullback-Leibler distance (Steyvers and Griffiths, 2013), Jensen-Shannon divergence (Boyack et al., 2011; Wagner et al., 2011) or Jaccard similarity (Agrawal et al., 2018). It is not uncommon that practitioners use a hybrid approach between the two methods, such as comparing distributions over words manually (or automatically) selected from the most relevant for each topic (Boyack et al., 2011; Chang et al., 2009; Greene et al., 2014; Yan et al., 2012).

In light of the fact that we can have these arbitrary transformations of the coordinate space, we require a different way of measuring agreement between models. We propose to evaluate the statistical robustness of a topic model via the extent to which it produces consistent estimates of pairwise document-document similarities. If perfectly robust, a model would produce the same similarity between two documents each time. An imperfect (yet useful) model will produce slight variations in each pairwise similarity score. Over many retrainings, it should converge to an asymptotic similarity value for each pair. Therefore, the use of inner pairwise comparisons eliminates the need for a *ground truth* or an *ad-hoc yardstick*, which is a critical concern in the literature (Velden et al., 2017). As we work in a vector space, we suggest using cosine similarity as a distance metric. Cosine similarity is a more intuitive interpretation of distance in the context of a spatial representation of knowledge (Garfield et al., 1978) than entropy-based metrics, with application beyond probabilistic topic models.

We expect that comparable models provide compatible representations of pairwise comparisons. Therefore, each document i, for each retraining k, is represented by the vector d_{ik}, where $k = 1, \dots, K$ and K is the total number of retrainings. We then compute the pairwise cosine similarity $s_{ijK} = \cos(d_{ik}, d_{jk})$ for each of the K retrainings (the only difference in the input being the random initialisation). Subsequently, we compute the average similarity $\bar{s}_{ijK}$ and the standard deviation σ_{ijK} associated. For each pair of documents, we have:

$$\bar{s}_{ij} = \frac{1}{K} \sum_{k}^{K} s_{ijk} \tag{1}$$

and

$$\sigma_{ij} = \left[\frac{1}{K} \sum_{k}^{K} \left[s_{ijk} - \bar{s}_{ij} \right]^2 \right]^{\frac{1}{2}} \tag{2}$$

In practice, for each training (k), we compute the similarity matrix M_k of all the pairwise similarities s_{ijk}. We finally calculate the average standard deviation across all the pairs. Fig. 1 illustrates a schematic representation of the experiment. Our generalised robustness metric is the asymptotic average standard deviation from all the unique ($i \neq j$ and $i < j$) pairwise similarities after K iterations:

$$\Phi_K = \frac{1}{C_{(N,2)}} \sum_{i}^{N} \sum_{j>i}^{N} \sigma_{ij} \tag{3}$$

where $C_{(N,2)} = \binom{N}{2}$ is the binomial coefficient of all the possible unique pairs with N documents.

The asymptotic standard deviation, Φ_K, is a comparison metric, which helps us determine the variation between retrainings of the pairwise similarities. Large values indicate more dispersion of the average similarity for each pair of documents. The more robust a topic model is to retrainings, the lower the value of Φ_K. There is not an absolute value that denotes goodness of fit (robustness) nor a general benchmark to compare to. Φ_K is a relative metric that depends on the data but can be used to determine which model introduces fewer spurious similarities.

On many occasions, practitioners will use topic models to infer the *most similar* documents in large corpora. The degree of relatedness in information retrieval or the distances for clustering applications, to name two examples, require that the higher similarity values be well defined. In order to reduce the effect of increased granularity and test for larger cosine similarity values, we extend the analysis to documents that, under at least one training, are highly similar. To do so, we filter out low similarities, and Eq. 3 takes now the following shape:

$$\Phi_K = \frac{1}{C^*} \sum_{i}^{N} \sum_{j>i}^{N} \sigma_{ij} \cdot \delta_{ij} \tag{4}$$

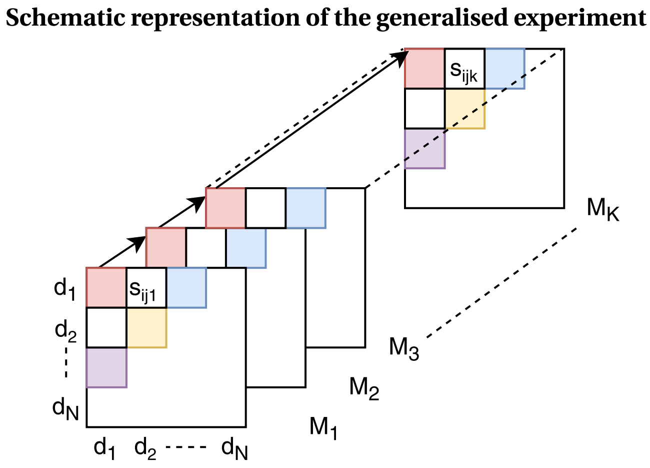


Fig. 1. We compute K similarity matrices M_k for each document pair. We then average the similarity for a given pair across the K retrainings and calculate the standard deviation for the pairwise similarity. Finally, we obtain the average standard deviation as an indicator of model stability.

where C^* is the number of observations left after filtering and:

$$\delta_{ij} = \begin{cases} 0 & \text{if} \quad s_{ijk} < \epsilon \quad \forall k \\ 1 & \text{if} \quad \exists s_{ijk} \geq \epsilon \end{cases} \tag{5}$$

where ϵ is the filtering value. That is, we calculate the robustness metric Φ_K taking into account only the pairwise similarities above ϵ (in at least one of the $k = 1, \ldots K$ similarity matrices, each for one retraining of the same model).

In addition to a potential greater relevance of "highly similar" documents for practitioner's objectives, we address a second non-trivial issue with the filters. As the number of dimensions grows, the degrees of freedom increase and the vectors representing each document in the latent space disperse. Therefore, there should be a larger concentration of close-to-zero similarities, which would, in turn, decrease the average standard deviation Φ_K. Put differently, as granularity increases, the changes in the retrainings will be absorbed along more dimensions, smoothing the dispersion in the pairwise similarities—i.e. lowering Φ_K. Furthermore, following each increment of the latent space, the distribution of pairwise similarities for any given document should systematically approach zero, thereby reducing the room for variation. Probabilistic models generate document-topic sparse matrices, whose sparsity will increase with the number of dimensions. This effect is not necessarily true for neural networks.

3.2. Descriptive power

It is common practice to base the latent space size selection on one's objectives and subsequent tasks. Often, higher dimensions are used for more granularity but become less useful for human interpretation (Chang et al., 2009). However, these parameters are also highly dependent on the size of the data and the variety inherently present in the text. Instead of selecting the number of topics, we suggest a way of comparing the descriptive power as a function of the number of topics. With this approach, we can confront retrainings of the same model (*i.e.* changing hyperparameters), different models (*e.g.* NMF *vs* LDA), and assess the information gained by including additional dimensions.

To get a handle on the explanatory power of Doc2Vec, LDA or NMF (or any topic modelling approach), we propose a straightforward procedure based on principal component analysis (PCA). We carry out the PCA decomposition on the document-topic vectors (researcher-topic vectors in this instance). The algorithm then reorders the principal components explained variance, and we plot their cumulative explained variance. This feature is present in the majority of standard distribution packages.

The PCA explained variance plot allows us to understand the extent to which each dimension allows differentiation among documents *vis-á-vis* the latent space. For example, a perfectly straight line running from the lower left to the upper right would indicate that each dimension contributes equally to explaining the variation among researchers— the increment of one unit in the dimension space adds a constant degree of information (variation). On the other hand, a curve that quickly reaches 1 (100%), perhaps after only $k < K$ dimensions, indicates that only those first k dimensions explain the variance. In other words, all dimensions beyond the first k do not add useful information. Thus, the explanatory power of a given topic model can be measured by the area over the curve (AOC) in these explained variance plots.

3.3. Reflect reality

Properly reflecting reality is, of course, the most important criterion for generating an abstract representation of data. It is often also the most difficult one to evaluate. Since the literature is filled with use-cases for probabilistic matrix factorisation models, we

Table 1
Corpus Statistics – Neuroscientists .

	Mean	StD	Min	Max
Publications/Document	120.5	79.1	50	1111
MeSH/Document	942.7	624.8	225	9985
MeSH Incidence	594.5	2877.5	1	114306

focus on validating the neural network approach that we are using in this manuscript. As the paper's locus lies in robustness and scalability, we do not analyse reality's reflection extensively. Here we limit ourselves to providing two small examples as evidence that, at the very least, the models' results do not directly oppose expectations, which would invalidate the results of the former sections. For this, we train document embedding vectors for different data and compare the output in different scenarios. In the Supplementary Material, Section 5 we validate the probabilistic methods.

4. Results

4.1. Robustness

For the robustness analysis, we focus on a subset of 13,936 researchers in the Neurosciences, sliced based on the heuristic journal classification discussed in Section 2.2 and described in detail in the Supplementary Material, Section 1. This choice is based purely on a desire to reduce the scale of the analysis so that all pairwise similarity scores can be calculated in a manageable amount of time and stored within the hardware at our disposal.[3] In this subsample, on average, documents have 940 non-unique MeSH from 120 distinct publications totalling over 1.6 million articles. Table 1 contains a summary of the corpus data.

We perform K = 100 retrainings for different sizes of the latent space (T = 10, 25, 50, 100, 250, 400, 800, 1000), for each of the three topic models. In order to better capture the potential systematic errors caused by the non-deterministic nature of the algorithms, we limit variability in the input as much as possible. First, the corpus is the same across all runs to enable comparison (Glaser et al., 2017; Klavans and Boyack, 2017; Velden et al., 2017), avoiding any deviation due to slightly different samples (Hecking and Leydesdorff, 2018). Second, not only data are the same, but the order in which they are fed to the algorithm in the training stage is the same (Agrawal et al., 2018). Finally, we use a fixed model tuning: all hyper-parameters are unchanged across runs with a fixed number of topics reducing the analysis of robustness to the stochastic initialisation (Belford et al., 2017). Following libraries best–practices, iterations have a learning decay upon convergence. Neural networks (Doc2Vec) run for ten epochs.[4] In other words: two different runs of the same model with the same latent space size only differ in the random seed. Thus, we expressly test for the variability that the practitioner cannot control (inherent to the non-deterministic nature of models). That is, we examine the solutions for the idiosyncratic effects of random initialisation, which, *a priori*, should not affect the internal coherence of topic models.

The results are summarised in Fig. 2. The top-left figure, corresponding to an $\epsilon = 0$ (no filtering-out of similarities), shows Φ_K from Eq. 3, while the other figures show Φ_K from Eq. 3 at different values of the filter ϵ (only pairs with similarity above ϵ are included). Overall, Doc2Vec provides robust and multi-purpose topic models that overcome the main difficulties described in the literature regarding stochastic matrix factorisation (LDA or NMF).

Fig. 2 allows us to grasp better the asymptotic behaviour of the two stochastic topic models and the neural-network approach. Without any filtering, we find evidence of a "breaking" point in both LDA and NMF, after which the model becomes significantly less robust. For the corpus in hand, for LDA, this happens for small-sized latent spaces. As the number of topics grows, the performance resembles that of Doc2Vec. For NMF, it evolves in the opposite direction. However, as we calculate the dispersion Φ_K including only the pairwise similarities larger than ϵ (under at least one retraining), the robustness of the stochastic topic models degenerates. Consistently, as ϵ increases, LDA and NMF show higher average dispersion of the pairwise similarities. Doc2vec, on the other hand, performs at a similar level independent of the filter. Complete results' figures with model-by-model asymptotic behaviour after filtering can be found in the Supplementary Material, Section 4.

In light of these results, one should be very careful at the tasks performed subsequent to topic modelling, particularly with stochastic dimensionality-reduction methods. In the particular case of retrieving ranks or lists of "most similar" documents, LDA and NMF display considerable variation. Therefore, it is plausible that the resulting documents are idiosyncratic to a specific training.

In the Supplementary Material, Section 5 we provide succinct evidence that the probabilistic models from our training would be considered robust under "traditional" entropy metrics. In the benefit of conciseness, we display results for LDA only (as discussed, the most extensively used topic modelling technique).

[3] The limitations arise in two stages: it is well known that LDA and NMF do not scale well in latent space size and number of documents. In addition, the pairwise similarity matrices M, which, in our case, for 14 thousand documents, are about 3 GB each. Therefore, it is computationally intensive to calculate the asymptotics of K (100 for our analysis) such matrices.

[4] Training time for NMF and LDA escalates exponentially with the latent space size while, for a similar number of iterations, Doc2Vec models do not require as much computational time. Using a machine with 30 cores, an average LDA model with a latent space size of 250 topics requires about 10 hours, with larger models using up to 20 and smaller requiring just minutes. For Doc2Vec, it requires about 2 hours, while the largest never exceeded 4 hours.

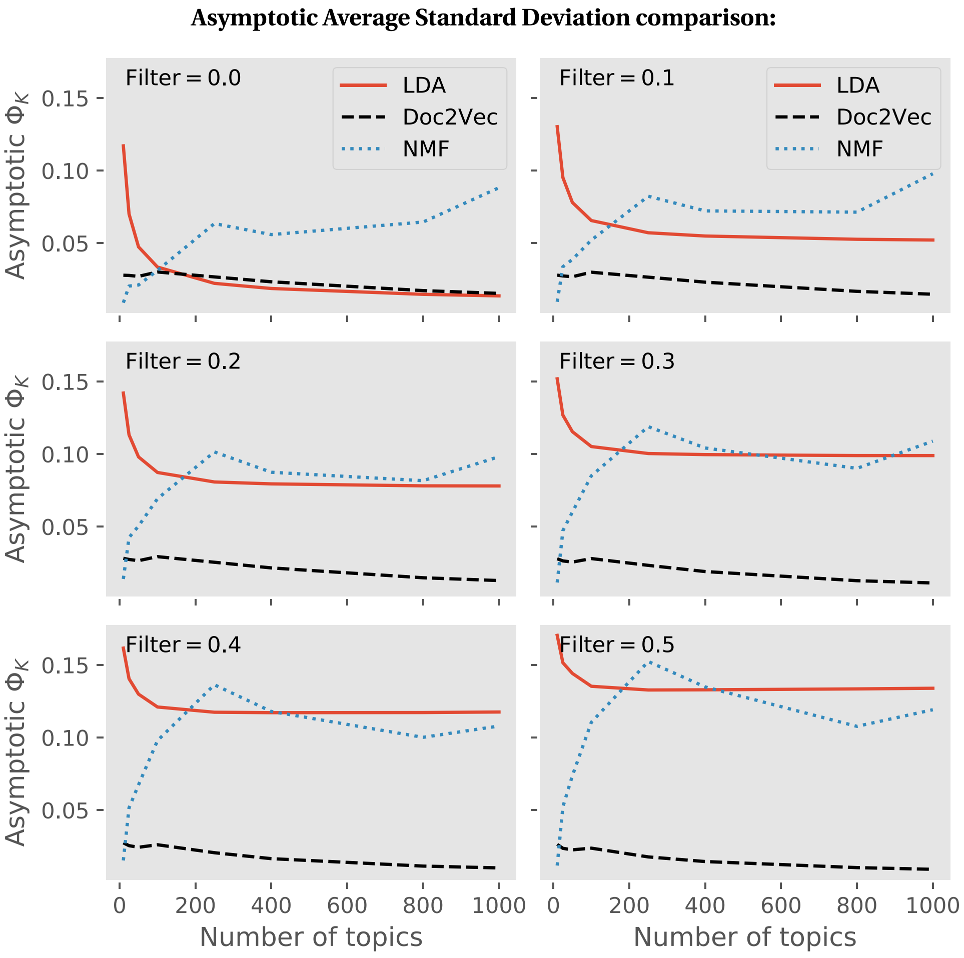


Fig. 2. Asymptotic value over multiple retrainings as a function of the number of topics. 10, 25, 50, 100, 250, 400, 800 and 1000 dimensions. Left-to-right and top-to-bottom, each figure displays Φ_K calculated after filtering out cosine similarities lower than $\epsilon = 0, 0.1, 0.2, 0.3, 0.4$ and 0.5 respectively.

4.2. *Descriptive power*

In order to study the scalability (explanatory power), we use the same models trained for the robustness analysis. Fig. 3 shows the explained variance of each dimension in a PCA-transformed space for multiple different topic sizes. The speed at which the different models converge to 1 (the slope of the lines) explains each additional dimension's incremental gain of information. For our particular corpus, albeit unevenly, both LDA and Doc2Vec carry descriptive power across the different dimensions under analysis. NMF, on the contrary, rapidly approaches 100 percent of the explained variance when trained with a large number of topics (250, 400, 800 and 100 in our analysis). NMF models lose descriptive power relative to LDA and Doc2Vec as the latent space size increases, eventually "breaking" in the models above 100 topics (for our data).

To discuss one particular example, for 250 topics we see that Doc2Vec is the closest to the diagonal, albeit far from overlapping. Still, a good 25 percent of the variance does reside in the last ~ 100 topics after the PCA transformation. On the other hand, we see that NMF models reach ~ 100 percent of variance explained within 100 dimensions, suggesting there is no real advantage of training models with a larger latent space for this data.

Visualising the PCA transformation does not single-handedly provide evidence of a *better* model. Instead, it provides a test for the marginal increase in explanatory power by expanding the latent space. In other words, we abstain from declaring Doc2Vec superior to LDA solely from a smaller area under the curve in the explained-variance-ratio plots. However, we argue that NMF models trained under this particular corpus provide no significant granularity gains for topic spaces larger than ~ 150 topics. In general, fine-tuning the model parameters decreases the area beneath the curves, attaining a significant gain in all the model dimensions. The models can scale in the number of dimensions carrying information at all times, allowing for a different optimisation depending on the objective of the training. PCA transformations allow us to measure and visualise whether there are such gains in scale.

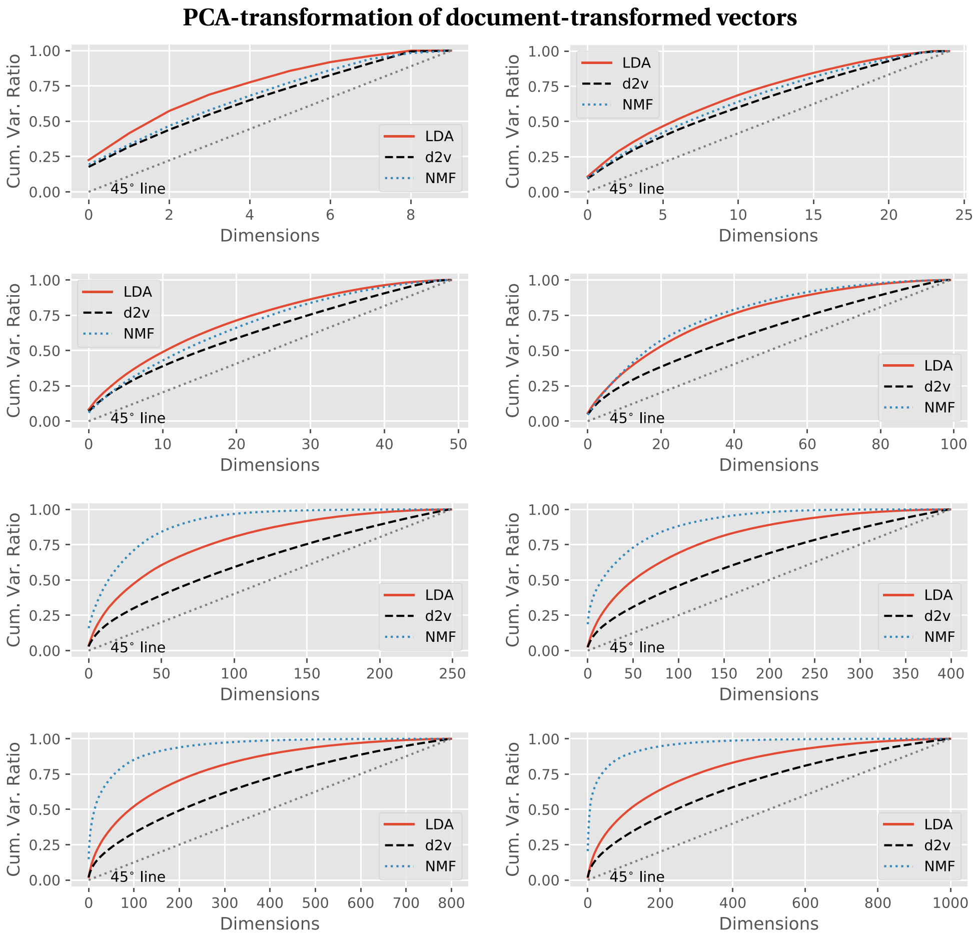


Fig. 3. Cumulative explained variance ratio of the transformed document (researcher) vectors, sorted by decreasing explained variance. The 45° line represents a model in which each dimension has the same descriptive power. The greater the Area Over the Curve (AOC), the greater the average descriptive power of each dimension.

4.3. *Concordance with reality*

Going forward, we are pursuing two main avenues of analysis for evaluating the extent to which document similarities produced by Doc2Vec reflect reality. For these tests, we will use different levels of data aggregation to represent the documents.

4.3.1. *Real-world communities*

First, we demonstrate the ability of document embeddings to generate communities that reflect their real-world associations. To this end, we train a single Doc2Vec model on the full corpus of 147,000 researchers from all biomedical sub-specialities. Upon training, the sub-field speciality labels are never fed into the models in any way.[5] In particular, we train for a latent space size of 400, and we project the document embeddings into two dimensions using t-SNE van der Maaten and Hinton (2008).[6] In Fig. 4 a (2D) t-SNE projection of the researcher embeddings is shown colour-coded by the sub-field speciality. From visual inspection, it is clear that, by and large, researchers of the same field cluster together, even if a 2D representation is limiting even though there are clearly some outliers and room for further investigation and/or refinement.

4.3.2. *Journal pairwise similarity*

Our second example, involves external information to identify pairs of documents that should be highly similar and valid on the Doc2Vec based similarity measures. Scholarly journals rarely shift topics suddenly over time. We exploit this feature and construct

[5] The speciality labels include, amongst others: Medicine, Molecular and Cell Biology, Dermatology, Radiology, Orthopedics, Dentistry or Obstetrics.

[6] t-SNE is a tool to visualise high-dimensional data. It converts similarities between data points to joint probabilities and tries to minimise the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

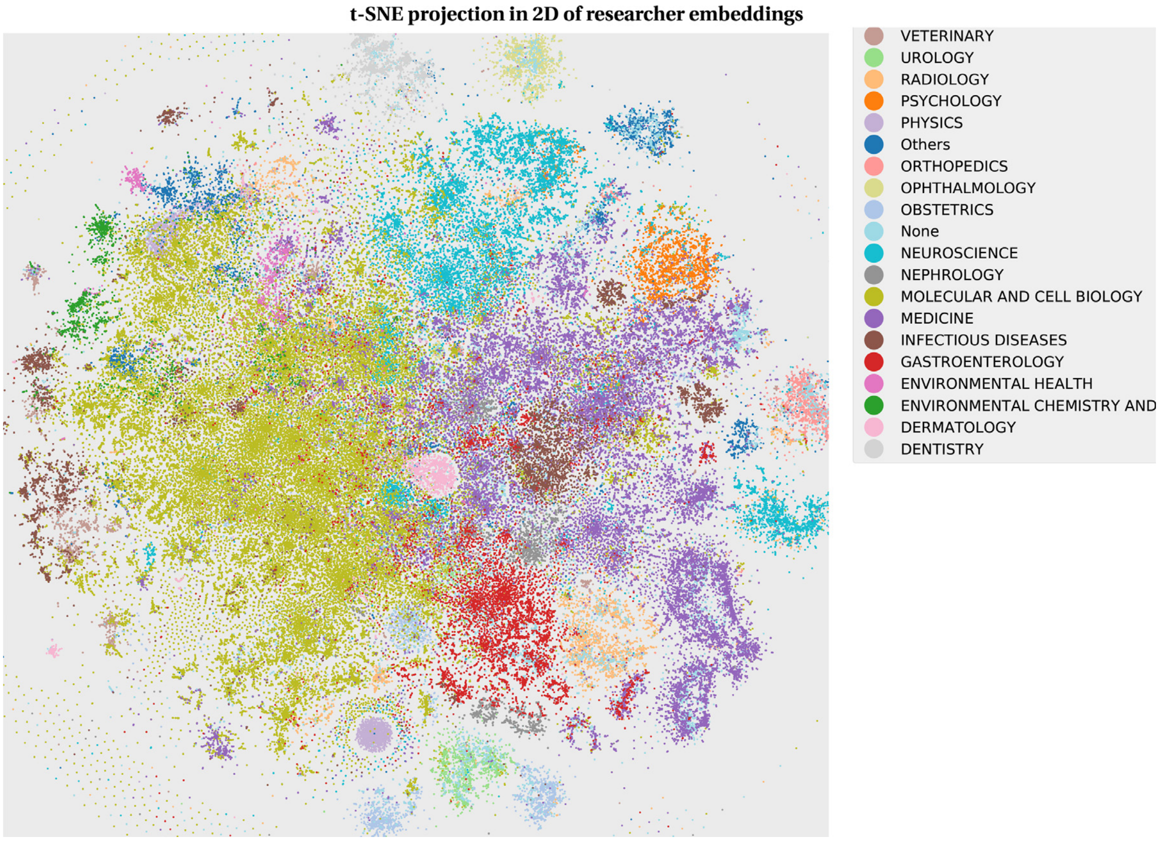
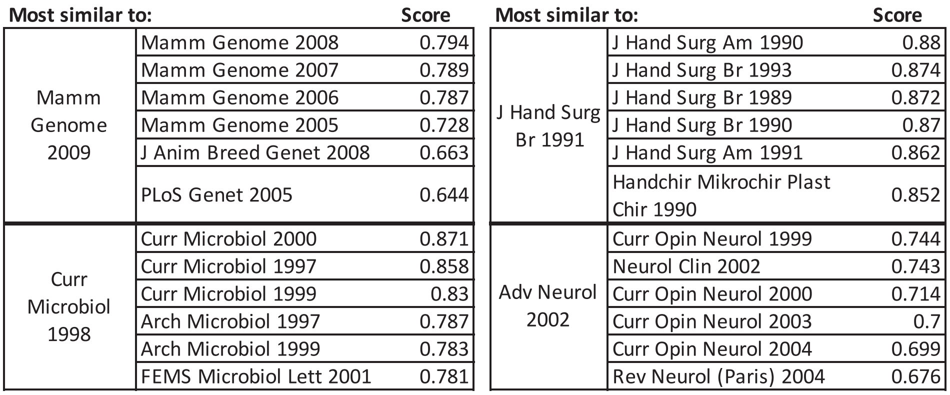


Fig. 4. Each point is a single researcher and colour indicates the researcher's field (the field in which the majority of his or her papers appeared).

Table 2
Top-6 documents by similarity and scores to a representative random sample. Trained with Doc2Vec using journal-year as documents.



Most similar to:		Score	Most similar to:		Score
Mamm Genome 2009	Mamm Genome 2008	0.794	J Hand Surg Br 1991	J Hand Surg Am 1990	0.88
	Mamm Genome 2007	0.789		J Hand Surg Br 1993	0.874
	Mamm Genome 2006	0.787		J Hand Surg Br 1989	0.872
	Mamm Genome 2005	0.728		J Hand Surg Br 1990	0.87
	J Anim Breed Genet 2008	0.663		J Hand Surg Am 1991	0.862
	PLoS Genet 2005	0.644		Handchir Mikrochir Plast Chir 1990	0.852
Curr Microbiol 1998	Curr Microbiol 2000	0.871	Adv Neurol 2002	Curr Opin Neurol 1999	0.744
	Curr Microbiol 1997	0.858		Neurol Clin 2002	0.743
	Curr Microbiol 1999	0.83		Curr Opin Neurol 2000	0.714
	Arch Microbiol 1997	0.787		Curr Opin Neurol 2003	0.7
	Arch Microbiol 1999	0.783		Curr Opin Neurol 2004	0.699
	FEMS Microbiol Lett 2001	0.781		Rev Neurol (Paris) 2004	0.676

a data set of Journal-Year documents using MeSH Terms as the document content. We characterise a Journal-Year document as the compilation of Medical Subject Headings (MeSH) published in a given periodical throughout a year. That is, we compile the MeSH terms for each article that appeared in the same journal during a year, grouping them in one single "document". That list of MeSH terms represents that Journal-Year.

We subsequently train a Doc2Vec topic model with Journal-Year documents ranging from 1985 to 2010 and generate inferred document embeddings (vectors) from 1985 to 2014 (4 years out-of-sample). The corpus comprises almost the entirety of our in-house PubMed database, with over 55,000 Journal-Year documents. We observe how the same journals in consecutive years show high similarity. In Table 2, we display the most similar journal-year documents and the similarity score to four randomly selected documents. As it can be seen, they are mostly similar to the same journals in close-by years. A 2D projection of these embeddings can be found in the Supplementary Material, Section 6 visually supporting the claim.

5. Discussion and limitations

In this manuscript, we have focused on comparing the performance of topic models from two different classes: dimensionality reduction and neural networks. As a representation of the first family, we have analysed one of the simplest, NMF, and the most used LDA. From the second family, we have provided evidence of document characterisation using Doc2Vec, a particular case of paragraph embeddings that arises from Word2Vec (the most extensively used word-embedding model). On a general level, there is an overlap in the representation of documents delivered by either approach (probabilistic or neural-network-based). A vector represents every document in a "knowledge space" and not as a discrete topics' classification. However, due to the sparse and positive components of the topic-term and document-topic matrices in topic models, LDA and NMF provide a more humanly interpretable output that enables classification tasks (topic labelling). On the other hand, neural network embeddings form a continuous space that is less readily explainable.

The comparison of models provided here is, to the best of our knowledge, the first effort to account for the idiosyncratic variability of solutions that escape the practitioner's analysis. We fix the data and hyperparameters for several retrainings, allowing only the stochastic initialisation to vary across runs. In order to compare their performance, we devised a robustness metric based on pairwise similarity. Furthermore, we do this for different levels of granularity. This system allows us to measure the extent to which each model would produce arbitrary relatedness (false positives) in the association (or disassociation) of documents.

Up until this point, we have provided evidence that neural-network-based topic models, in particular Doc2Vec, provide a more reliable—replicable—characterisation of documents. In turn, Doc2Vec can scale (increase the number of topics) without losing explanatory power and remains within manageable computing times even for personal computers. Additionally, we provide evidence of different levels of support of the three models for different ranges of similarities— *i.e.* we show that for LDA and NMF, robustness decreases as the pairwise similarity increases. Our results suggest that solutions obtained with stochastic dimensionality reduction methods are, on many occasions, contingent on a particular training. Even when traditional coherence metrics and ground-truth tests support the models, the optimality is frequently incidental. However, this analysis precludes us from claiming the superiority of one particular solution over the others. One would ideally study the complete set of hyper-parameters and data-sample combinations to determine the resolution of a solution. Instead, our analysis provides practitioners with additional tools to evaluate their particular applications. Additionally, we provide a quantitative measure to the well-known issue of instability of probabilistic topic models.

While we take the analysis to the asymptotic extreme by using complete (all-*vs*-all) pairwise similarities of 100 models to estimate robustness, practitioners can reject a model's performance with fewer trainings and comparisons. As we show, the instability is larger as similarity increases so, comparing a subsample of *a priori* similar documents amongst 3–5 retrainings should suffice to discard a model's robustness.

These observations raise one question: should we then turn our backs on probabilistic models favouring document embeddings? Not necessarily. First, neural network methods are rather data-thirsty. That is, for a good embedding characterisation, one needs large amounts of data (Mikolov et al., 2013). Under these circumstances, they outperform probabilistic methods both in computational speed and robustness. However, for small document-term matrices, convergence to a workable solution is rarely achieved. In this case, dimensionality reduction techniques (probabilistic, such as NMF, LDA or pLSI) or factor analysis (such as PCA) might be more suitable and accurate. Second, neural network methods do not provide humanly readable topics. In contrast to "traditional" topic models, there is no *top-words* output that makes neural network embeddings intelligible. Thus, additional work is required for classification or labelling tasks, over-complicating a task that simpler methods still effectively deliver. Proper understanding of the data and task requirements, along with the properties of each model, is more likely to provide the best results.

More work is needed to prove concordance with reality of neural network embeddings. Many of these solutions are still very recent and need more testing before they gain momentum in the scientometrics community and the social sciences in general. We have provided two rough examples that should provide a sanity check for the conclusions extracted in the rest of the analysis. We show how they can work in different contexts, at various aggregation levels and for varied tasks. Ultimately, the quality and performance of these tasks improve with parameter tuning, which we have overlooked in this work.

The explicit omission of parameter tuning is not the only limitation of the analysis above. As with any empirical study, we cannot discard any biases in the analysis. First, we put the models to the test under a very particular dataset comprised of MeSH terms. While that choice comes with several advantages, those may be offset by a lack of comparison with other samples. Furthermore, it raises questions concerning to what extent the results are particular to the characteristics of the corpus. Second, we train the model on relatively large documents (943 terms on average). Paragraph embeddings tend to overfit when documents are too short (Ai et al., 2016). Our corpus is in a sweet spot where probabilistic topic models can still be applied, and neural-network document embeddings have enough data to converge in the training stage. For a smaller corpus, training neural networks becomes more challenging, which compromises the conclusions of this work. We could see a potentially better performance from traditional topic modelling methods. For shorter documents, however, it has been shown that word embedding aggregations (Word2Vec) can represent short paragraphs adequately (De Boom et al., 2016).

The choice of models in this work illustrates the most accessible (present in most software packages) and widely used topic modelling techniques. Nevertheless, we have left out other commonly used models and have demonstrated the ability to overcome some of the problems discussed above. Amongst the "traditional" topic models, there have been numerous developments which we have not tested here, such as Pairwise-Mutual Information (PMI), probabilistic Latent Semantic Indexing (pLSI), Latent Semantic

Analysis (LSA). Many hybrid methods have been developed to complement probabilistic topic models in search of stability (Agrawal et al., 2018; Belford et al., 2017; Velden et al., 2017). On the other hand, there are other models amongst the "new breed" of neural-network-based models. FastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014) and Word2Vec (from where Doc2Vec stems) are other word (and sub-word) embedding models of the same generation (log-bilinear prediction-based semi-supervised models that generate static embeddings). See the Supplementary Material, Section 7 for an extended analysis of Word2Vec.

Finally, an important limitation to the analysis proposed above lies in the lack of a benchmark data set that provides a comparison of topic extraction results. Nevertheless, the evolution of research involving prior topic modelling techniques suggests that reflection of reality tasks will be amongst the first to be tested by the community. While still a nascent area of interest, the work by Banerjee et al. (2018) provides a first successful (contrasted) effort in using word embeddings for information categorisation tasks; Ai et al. (2016) employ paragraph embeddings for information retrieval in short texts; and Thijs (2019) provides evidence of Doc2Vec-generated similarity between scholarly article sections, following *a priori* rational expectations. In a document retrieval exercise, Dai et al. (2015) show evidence of the superiority of paragraph embeddings over LDA and a static bag-of-words (no topic modelling) approach on a set of arXiv publications.

Beyond these simple use cases, embedding approaches can also be used to accomplish more intricate tasks, and practitioners have started to apply these methods in their publications. Lenz et al. (2020) show how to measure the diffusion of innovations through paragraph embeddings by studying the dynamic properties of the knowledge space. Following a similar analysis, it is possible to characterise *ex-post* novelty, hot topics and disappearing topics through topic models (Ballester and Penner, 2019). On the other hand, Ayoubi et al. (2020) use embeddings to measure the disparity between past and present grant applications of researchers.

Exciting avenues of future research emerge from the understanding of word and document embeddings. Using these methods, it would be possible, for instance, to characterise the distancing process between a researcher and her mentor. Or the knowledge fit between peers in mobility events. With adequate data and careful training, it might be possible to subtract (or add) the contribution of, for example, a co-author to a publication. Similarly, if researchers can disentangle the direction of methodologies, one could potentially divide empirical and theoretical contributions in a publication.

CRediT authorship contribution statement

Omar Ballester: Methodology, Formal analysis, Software, Writing – original draft, Visualization, Writing – review & editing. **Orion Penner:** Conceptualization, Funding acquisition, Supervision, Resources, Data curation, Writing – review & editing.

Acknowledgements

Preliminary versions of this manuscript have been presented at the Global Tech Mining Conference (Atlanta, 2019), the IV Summer School in Science and Technology Indicators (KUL-EPO, Vienna, 2019), International Conference on Scientometrics and Informetrics (Rome, 2019) and EPFL Workshop on Computational Methods in Social Science (2019). We wish to thank the comments and remarks received from referees and attendants of those events, who have played a role in the improvement of the analysis presented above. A working-paper version of this work was published as part of the ISSI 2019 proceedings (Ballester & Penner, 2019). This work was supported by the Swiss National Science Foundation via project number 167943. We would like to thank Martin Rosvall and Jevin West for providing a structured version of the journal classifications used in Eigenfactor. This work would not have happened without the support of Gaétan de Rassenfosse. Finally, we would like to thank two anonymous comments by reviewers who helped improve the final version of the paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.joi.2021.101224

References

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology, 98*, 74–88. 10.1016/j.infsof.2018.02.005.

Ai, Q., Yang, L., Guo, J., & Croft, W. B. (2016). Analysis of the paragraph vector model for information retrieval. In *Proceedings of the 2016 ACM international conference on the theory of information retrieval*. In *ICTIR'16* (pp. 133—142). New York, NY, USA: Association for Computing Machinery. 10.1145/2970398.2970409.

Ayoubi, C., Barbosu, S., Pezzoni, M., & Visentin, F. (2020). What matters in funding: The value of research coherence and alignment in evaluators' decisions. *MERIT Working Papers*. United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT). https://ideas.repec.org/p/unm/unumer/2020010.html.

Ballester, O., & Penner, O. (2019). Evolution of Topics and Novelty in Science. In *Proceedings of the 17th International Conference on Scientometrics and Informetrics Vol II*. In *ISSI 2019* (pp. 1606–1611). Rome, Italy: International Society of Scientometrics & Informetrics.

Banerjee, I., Chen, M. C., Lungren, M. P., & Rubin, D. L. (2018). Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *Journal of Biomedical Informatics, 77*, 11–20. 10.1016/j.jbi.2017.11.012.

Belford, M., Namee, B. M., & Greene, D. (2017). Stability of topic modeling via matrix factorization. *CoRR, abs/1702.07186* arXiv:1702.07186.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The eigenfactor^{TM} metrics. *Journal of Neuroscience, 28*(45), 11433–11434. 10.1523/JNEUROSCI.0003-08.2008.

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine, 27*(6), 55–65. 10.1109/MSP.2010.938079.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research: JMLR, 3*, 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Borner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*(1), 179–255. 10.1002/aris.1440370106.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Borner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS one, 6*(3), 1–11. 10.1371/journal.pone.0018029.

Braam, R., Moed, H., & Raan, T. (1991). Mapping of science by combined co-citation and word analysis. i. structural aspects. *JASIS, 42*, 233–251. 10.1002/(SICI)1097-4571(199105)42:43.0.CO;2-I.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155–205. 10.1007/BF02019280.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd international conference on neural information processing systems*. In *NIPS'09* (pp. 288–296). USA: Curran Associates Inc.. http://dl.acm.org/citation.cfm?id=2984093.2984126.

Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. *CoRR* arXiv:1507.07998.

De Boom, C., Canneyt, S. V., Demeester, T., & Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters, 80*, 150–156. 10.1016/j.patrec.2016.06.012.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science (New York, N.Y.), 122*(3159), 108–111. 10.1126/science.122.3159.108.

Garfield, E., Malin, M. V., & Small, H. R. (1978). Citation data as science indicators (pp. 179–208).

Glaser, J., Glanzel, W., & Scharnhorst, A. (2017). Same data—different results? towards a comparative approach to the identification of thematic structures in science. *Scientometrics, 111*(2), 979. 10.1007/s11192-017-2295-0.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management, 41*(6), 1548—1572.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? stability analysis for topic models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Machine learning and knowledge discovery in databases* (pp. 498–513). Berlin, Heidelberg: Springer Berlin Heidelberg.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235. 10.1073/pnas.0307752101.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*. In *EMNLP '08* (pp. 363–371). Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1613715.1613763.

Hecking, T., & Leydesdorff, L. (2018). Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984–998. 10.1002/asi.23734.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR* arXiv:arXiv:1405.4053..

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*, 788–791. 10.1038/44565.

Lenz, D., & Winker, P. (2020). Measuring the diffusion of innovations with paragraph vector topic models. *PloS one, 15*(1), 1–18. 10.1371/journal.pone.0226685.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics, 3*, 211–225. 10.1162/tacl_a_00134.

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology, 63*(10), 1973–1986. 10.1002/asi.22628.

van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *The Journal of Machine Learning Research*.

Mei, Q., Shen, X., & Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*. In *KDD'07* (pp. 490—499). New York, NY, USA: Association for Computing Machinery. 10.1145/1281192.1281246.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR* arXiv:1301.3781.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. In *HLT '10* (pp. 100–108). Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1857999.1858011.

Noyons, E. C. M., & van Raan, A. F. J. (1994). Bibliometric cartography of scientific and technological developments of an r & d field. *Scientometrics, 30*(1), 157–173. 10.1007/BF02017220.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). http://www.aclweb.org/anthology/D14-1162.

Řehuřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems, 28*(1), 4:1–4:38. 10.1145/1658377.1658381.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118–1123. 10.1073/pnas.0706851105.

Steyvers, M., & Griffiths, T. (2013). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Psychology Press.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 990–998).

Thijs, B. (2019). *Paragraph-based intra- and inter- document similarity using neural vector paragraph embeddings* (pp. 1900–1911). ISSI. https://lirias.kuleuven.be/retrieve/548804DISSI2019_WordEmbeddings_Proceedings.pdf.

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3*(3), 11:1–11:29. 10.1145/1552303.1552304.

Velden, T., Boyack, K. W., Glaser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics, 111*(2), 1169–1221. 10.1007/s11192-017-2306-1.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*(1), 14–26. 10.1016/j.joi.2010.06.004.

Yan, E., Ding, Y., Milojević, S., & Sugimoto, C. R. (2012). Topics in dynamic research communities: An exploratory study for the field of information retrieval. *Journal of Informetrics, 6*(1), 140–153. 10.1016/j.joi.2011.10.001.