When less is more?

Topic model performance across title, abstract, and full-text corpora

Francis Lareau and Christophe Malaterre

1. Introduction

Topic modeling has emerged as an indispensable computational tool for analyzing large volumes of scientific literature, particularly in fields such as history, philosophy, and sociology of science. By automatically identifying thematic patterns within textual data, topic modeling enables researchers to track research trends, identify paradigm shifts, analyze the evolution of scientific vocabulary, and map the interrelations of scientific themes. However, a crucial methodological question remains largely unexplored in the case of article-based corpora: what level of textual granularity—titles, abstracts, or full texts—provides the optimal balance between computational efficiency and analytical depth?

This question has significant practical implications for mining corpora of academic articles. Full-text analysis demands considerable computational resources for acquisition, preprocessing, and analysis, while title or abstract analysis offers a more economical alternative. Yet, the potential information loss when using only titles or abstracts must be weighed against these efficiency gains. Despite the importance of this methodological consideration, systematic comparisons across text structures remain scarce in the literature.

This contribution examines the performance of two prominent topic modeling approaches—Latent Dirichlet Allocation (LDA) and BERTopic—across three text structures (titles, abstracts, and full texts). Through a set of quantitative evaluations, we seek to determine whether the additional resources required for full-text analysis are justified by proportional analytical gains, or if more parsimonious approaches using titles or abstracts might suffice for certain research objectives.

2. Methodological considerations

The comparison centers on two fundamentally different topic modeling approaches: LDA, a classical statistical technique based on word-count vectors and Dirichlet distributions (Blei et al., 2003), and BERTopic, a more recent approach leveraging large language model embeddings (Grootendorst, 2022).

LDA represents documents as bags-of-words and models topics as latent variables following a Dirichlet distribution. Its relatively straightforward implementation has made it the de facto standard in computational text analysis for over a decade. LDA can handle documents of varying lengths, making it suitable for all three text structures under investigation.

BERTopic, conversely, represents a newer generation of topic modeling that capitalizes on contextual embeddings from large language models. Rather than treating topics as latent variables, BERTopic identifies topics through document clustering in the embedding space. Until recently, BERTopic was limited in its ability to handle long texts due to token limitations in transformer models. However, recent innovations in long-text embeddings have expanded BERTopic's applicability to full-text corpora, making this comparative analysis particularly timely.

This study used a corpus of 3,698 full-text articles from three leading astrobiology journals (1974–2023), curated by Malaterre and Lareau (2023). Only documents that include a title, abstract, and full-text were retained, totaling 3,542 documents. After standard preprocessing—tokenization, POS tagging, and lemmatization— stopwords were excluded (Manning, Raghavan and Schütze, 2008). Only nouns, verbs, adverbs, and adjectives were kept, and documents were vectorized into term-document matrices for LDA and BERTopic. LDA modeling was performed using a Python API with a word frequency TDM. For BERTopic, document embeddings were generated from full texts using the high-scoring stella model (stella_en_1.5B_v5) based on Alibaba-NLP (Zhang et al., 2025). The BERTopic pipeline was executed via Python API, utilizing the TDM for word ranking and outlier reassignment. Note that the same protocol was used for titles and abstracts.

3. Comparative analysis framework

To systematically compare model performance across text structures, a multi-faceted evaluation framework is essential. Here, as a start, we examine performance with three complementary metrics:

Topic Diversity: Assesses whether topics within a model are represented by distinct word sets. Since this is the percentage of distinct words among all the top words representing topics, greater diversity suggests more differentiated topics with less semantic overlap.

Joint Recall (mJIR): Measures how effectively a topic's representative words collectively recall the documents assigned to that topic (Lamirel et al., 2025). Higher recall indicates better alignment between topic keywords and document content:

$$mJIR = \frac{1}{|D|} \sum_{c=1}^{K} |\{d \in c | \exists i, i \in Top_W[c] | d[i] \neq 0\}|$$

where W is the number of top words chosen as description of any cluster c, $|D|$ is the number of documents in the corpus, $Top_W[c]$ is the set of the top W words describing topic c, $d[i]$ represents the presence/absence of word i in the document d.

Topic Coherence: Evaluates whether the words describing each topic semantically belong together in a meaningful way. Higher coherence suggests more interpretable topics. We used the coherence CV of Röder et al. (2015).

As a complementary measure, we also calculated the standard deviation of document counts per topic to capture how evenly documents are distributed. Lower values indicate balanced topic representation, while higher values suggest dominance by a few topics. For LDA, although each document has a full topic distribution, we assigned documents to their dominant topic to enable comparability with crisp clustering models such as BERTopic.

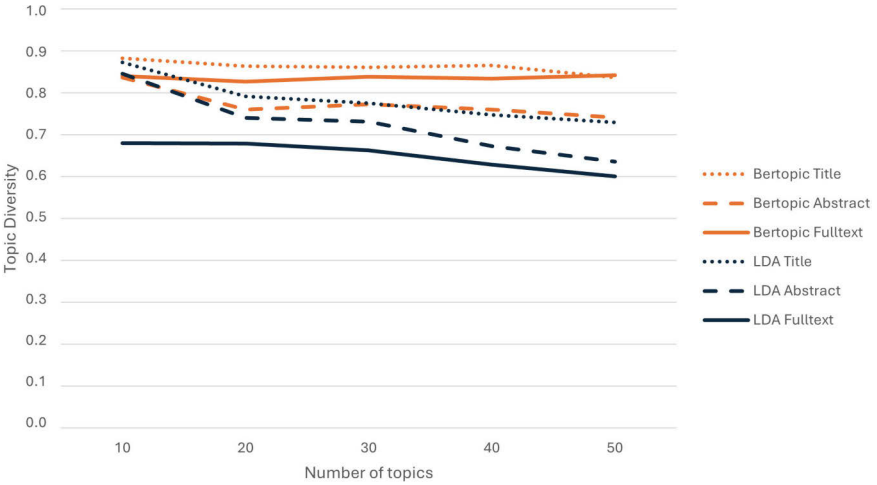
These metrics, when considered collectively, provide a first framework for evaluating model performance across different text structures and algorithm choices. Of course, the models should also be evaluated along other dimensions, notably qualitatively (in terms of topic interpretability) but also in how document assignment is balanced across topics. This is work under progress.

4. Preliminary results

The framework consisting of the different metrics was applied to 5 models of approximately 10, 20, 30, 40 and 50 topics, and this for each type of model, either BERTopic or LDA trained on titles, abstracts or full text:

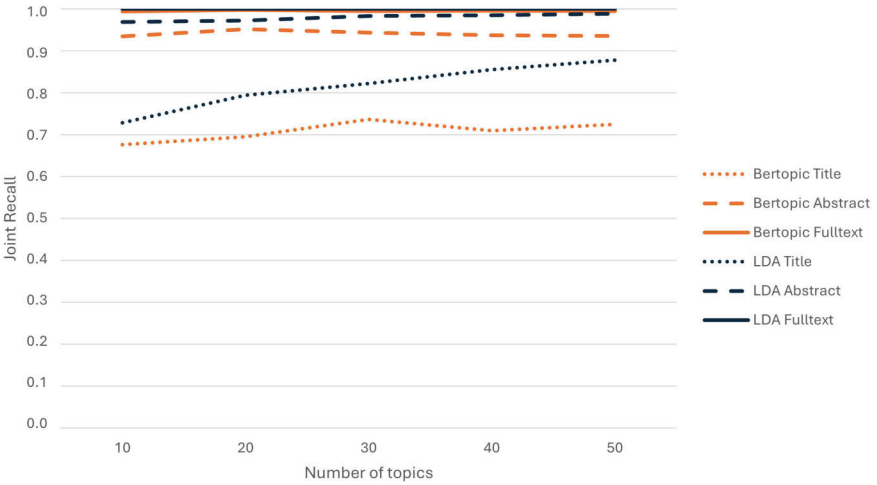
Topic Diversity: Preliminary diversity results show that all models offer a minimum diversity of at least 60%. The best-performing models being BERTtopic trained on titles or full text, offering a diversity of more than 83% (Fig. 1).

Fig. 1: Topic diversity performance comparisons between topic models (10 top-words by topic).



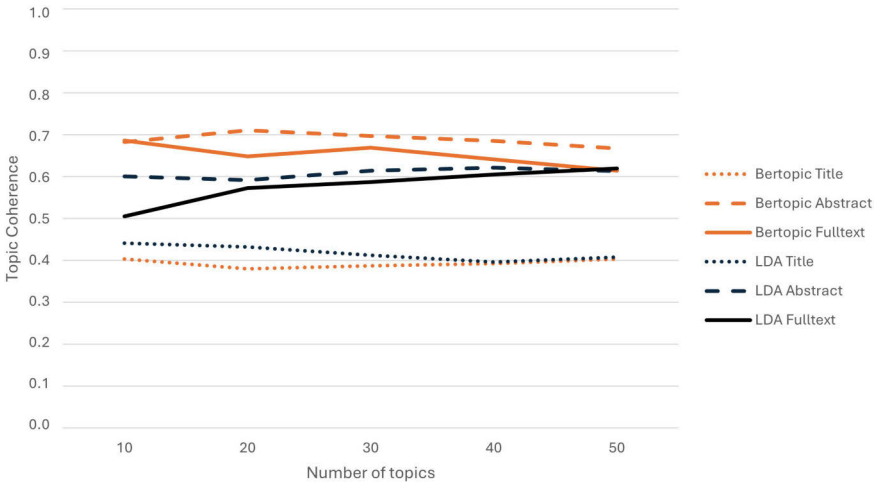
Joint Recall: The preliminary results of joint recall show two clear trends: that LDA models outperform BERTopic models when considering the same document type, and that full text outperforms abstracts, which in turn outperform titles, when considering the same type of algorithm (Fig. 2).

Fig. 2: Joint recall performance comparisons between topic models (10 top-words by topic).



Topic Coherence: Preliminary coherence results show that BERTopic models trained on abstracts and full text perform best, followed by LDA models on the same document structures. Models trained on titles perform worst (Fig. 3).

Fig. 3: Coherence CV performance comparisons between topic models (10 top-words by topic).



Standard deviation (SD): The dispersion of the documents in the topics varied substantially by model and document type when comparing the *N*=25 topics models. LDA applied to full texts yielded a standard deviation of approximately 52.2, closely aligned with the value obtained from abstracts (SD ≈ 53.4), suggesting; by contrast, LDA on titles produced the lowest dispersion (SD ≈ 28.4). In comparison, BERTopic exhibited markedly higher variability: full texts produced the greatest spread (SD ≈ 116.9), abstracts showed intermediate dispersion (SD ≈ 57.0), and titles yielded unexpectedly high variability (SD ≈ 96.8). This indicates that LDA produces more homogeneous document allocations to topics regardless of text length, and more so than BERTopic which tends to generate less balanced topics in terms of number of documents assigned to them.

5. Key findings

5.1 Title-based models: information scarcity challenge

Topic models based solely on titles consistently underperform across most metrics. Although the models trained on the titles tend to generate topics with high diversity, especially the BERTopic model (Fig. 1), they significantly lag behind their abstract and full-text counterparts for the other two metrics, namely topic coherence and most notably document recall (Fig. 2 and 3). This means that title-based models do not manage to properly cover the entire spectrum of documents.

The primary challenge appears to be information scarcity. Titles, by design, provide only the most succinct representation of document content, often emphasizing novelty or appeal over comprehensive thematic coverage. This limitation manifests in poor coherence scores and problematic document-topic assignments. However, it is noteworthy that despite these limitations, BERTopic title models can still identify topics scoring high on diversity and covering about 70 to 80% of documents (as measured by recall).

5.2 Abstract models: the Goldilocks zone

Abstract-based models consistently demonstrate strong performance across evaluation metrics. The BERTopic abstract model, in particular, achieves the highest coherence scores (Fig. 3), though its performance in terms of diversity and recall is not as good as the BERTopic full-text model (Fig. 1 and 2). LDA abstracts models consistently offer better coherence and diversity scores compared to their full-text counterparts, though again recall is less satisfactory.

This overall excellent performance may be attributed to the nature of abstracts themselves. As condensed summaries designed to highlight key findings and themes, abstracts offer a focused representation of document content without the noise and methodological details often present in full texts. This balance between information density and signal clarity positions abstracts in a "Goldilocks zone" for topic modeling—providing sufficient information for robust topic identification without the computational and analytical complexities of full-text processing.

5.3 Full-text models: information overload effects

Full-text models demonstrate complex performance patterns. While they achieve strong joint recall scores (Fig. 2)—indicating excellent coverage of document content by topic keywords—they exhibit lower performance than their abstract-based counterparts, especially in terms of topic coherence (for both LDA and BERTopic), but also diversity (for LDA, though not BERTopic).

One possible explanation is the presence of methodological aspects alongside conceptual content in full-text documents: this may indeed blur both the consistency of topics by mixing methodological terms with substantive ones (thereby lowering topic coherence measures) and their diversity (at least in the case of the LDA model), possibly by adding cross-thematic methodological terms. This may suggest an "information overload" effect, where the wealth of details in full texts can complicate the identification of clear thematic boundaries. However, full-text models may be useful to capture nuanced themes that might be absent from title- and abstract-based models, particularly regarding methodological approaches and secondary research themes. Such qualitative aspects need to be further examined.

6. Practical implications

These findings have several practical implications for researchers employing topic modeling in scientific literature analysis:

Research-Goal Alignment: The choice of text structure should align with specific research objectives. For broad thematic mapping, abstract models offer an excellent balance of performance and efficiency. For fine-grained analysis of methodological approaches or secondary themes, full-text models may provide crucial insights despite their computational costs.

Resource Optimization: The strong performance of abstract models suggests that in many cases, researchers can achieve robust results without the substantial computational and preprocessing requirements of full-text analysis. This is particularly valuable for large-scale studies or when working with limited computational resources.

Multi-Level Modeling: Rather than viewing text structures as competing alternatives, researchers might benefit from a multi-level approach that leverages the complementary strengths of different text structures. For instance, abstract models might serve for initial thematic mapping, with full-text analysis reserved for deeper exploration of specific areas of interest.

Model Selection: BERTopic generally outperforms LDA across text structures, particularly for coherence and diversity metrics (though not recall nor standard deviation). This suggests that despite their greater computational requirements, embedding-based approaches offer tangible analytical benefits that may justify their adoption.

Conclusion

The question of whether less is more in topic modeling depends fundamentally on research objectives and resource constraints. Abstract-based models emerge as particularly strong performers, offering a compelling middle ground that captures most major thematic structures without the computational burden of full-text analysis. Title-based models, while computationally lightweight, sacrifice too much information for most analytical purposes. In any case, the different models need to be submitted to further complementary evaluations, notably in terms of topic interpretability (for instance, by examining top-words and top-documents assigned to the models, depending on textual structure), but also in terms of how document assignment is representatively balanced across topics (notably to avoid large generalist topics as well as extremely small overfitted topics). These are aspects that we are currently investigating.

These findings suggest that the field might benefit from developing hybrid approaches that intelligently combine information from different text structures, potentially extracting the focused signal from abstracts while selectively incorporating the richer detail available in full texts. As topic modeling continues to evolve with advances in language models and computational techniques, such multi-level approaches may represent the next frontier in scientific literature analysis.[1]

1 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Acknowledgments

F.L. acknowledges funding from Canada Social Sciences and Humanities Research Council (Postdoctoral Fellowships 756–2024-0557, Grant 430–2018-00899). C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430–2018-00899) and Canada Research Chairs (CRC-950-230795). The authors thank Arno Simons, Adrian Wüthrich, Michael Zichert and Gerd Graßhoff for organizing the *LLMs in HPS* workshop (TU Berlin, 2025) and inviting them to contribute to this volume.

References

Blei, D M, Ng, A Y and Jordan, M I (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan): 993–1022.

Grootendorst, M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint* arXiv:2203.05794.

Lamirel, J-C, Lareau, F and Malaterre, C (2025) Old but Not Obsolete: Bag-of-Words vs. Embeddings in Topic Modeling. *Proceedings of the 20th International Conference on Scientometrics and Informetrics*, Jun 2025, Yerevan, Armenia.

Malaterre, C and Lareau, F (2023) The Emergence of Astrobiology: A Topic-Modeling Perspective. *Astrobiology*, 23(5): 496–512. https://doi.org/10.1089/ast.2022.0122

Manning, C D, Raghavan, P and Schütze, H (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Röder, M, Both, A, and Hinneburg, A (2015) Exploring the Space of Topic Coherence Measures, *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. New York: ACM, 399–408.

Zhang, D, Li, J, Zeng, Z and Fulong, W (2025) Jasper and Stella: distillation of SOTA embedding models. *arXiv preprint* arXiv:2412.19048.