Which contextual topic modelling algorithm is best?

Darren Craig Roos[1][0000−0002−0405−524X] and Katherine Mary Malan[1][0000−0002−6070−2632]

Department of Decision Sciences, University of South Africa, Pretoria, South Africa
21001154@mylife.unisa.ac.za

Abstract. The question of which contextual topic modelling algorithm performs best has become increasingly important as the field rapidly develops new approaches. However, existing evaluations typically focus on limited datasets and metrics, often claiming superiority for novel algorithms. This study presents a comprehensive empirical evaluation of eleven contextual topic modelling algorithms across ten diverse datasets, five numbers of topics, and four performance metrics, resulting in 22,000 metric evaluations. Rather than identifying a single superior algorithm, our results reveal clear evidence of performance complementarity: different algorithms excel on different problem instances and under different evaluation criteria. Through aggregate performance analysis, pairwise dominance comparisons, and multi-objective Pareto frontier analysis, we demonstrate that algorithmic dominance varies significantly across problem instances. Most remarkably, in 84% of cases, all algorithms are Pareto optimal when considering all metrics simultaneously, indicating that each offers unique strengths that cannot be dominated by others. These findings challenge the common practice of claiming algorithmic superiority and suggest that algorithm selection should be guided by specific problem characteristics and performance priorities rather than blanket recommendations. Our work contributes to the growing recognition that performance complementarity is fundamental to computational problems, extending this concept to contextual topic modelling and providing a foundation for future algorithm selection research. Code used to conduct this study is provided. [1]

Keywords: Natural language processing · Topic modelling · Performance analysis · Performance complementarity.

[1] https://github.com/AlgorithmicAmoeba/tm_framework

1 Introduction

When working with a large corpus of documents, it is often useful to be able to understand common themes that occur within the corpus. Topic modelling algorithms are designed to solve this exact problem, offering unsupervised machine learning approaches that process corpora and identify common themes—called topics. These algorithms represent topics as collections of words that frequently co-occur in documents. For example, words such as "climate", "warming", "emissions", and "greenhouse" might collectively represent the topic of climate change, while terms like "algorithm", "complexity", "computation", and "efficiency" might correspond to computational theory. Topic modelling has been applied to various fields including bioinformatics [10,11], medicine [14], law [41] and text summarisation [16,21,36].

The landscape of topic modelling has evolved significantly over the past two decades. Traditional probabilistic approaches like Latent Dirichlet Allocation (LDA) [6,7] modelled documents as mixtures of topics, where each topic represented a probability distribution over words. Concurrently, matrix factorisation methods such as Non-negative Matrix Factorisation (NMF) [31] approached the problem by decomposing term-document matrices into lower-dimensional representations. Despite their theoretical elegance and widespread adoption, these classical methods suffered from limitations in capturing semantic relationships beyond word co-occurrence statistics. The emergence of neural topic models [34,48] addressed some of these shortcomings by leveraging neural networks to learn more expressive document representations.

This evolution culminated in Contextual Topic Modelling (CTM) algorithms [3,4]. They represent a subclass of topic models that have emerged from the recent Large Language Model (LLM) revolution [15,42,51,52]. These models use information, typically in the form of embeddings from LLMs, to enhance their performance. Such embeddings have been demonstrated to encapsulate a plethora of semantic information [9,19,45], providing contextual understanding that traditional topic models lack. This additional semantic richness enables CTMs to better capture nuanced relationships between words and documents, potentially leading to more coherent and interpretable topics.

Recent literature reveals numerous CTM approaches [2,27,46,53,54]. Typically when a new CTM is introduced in literature, a publication will outline how the novel algorithm works. This is usually followed by an empirical investigation into its performance on selected datasets, comparing it to existing approaches. Almost always, the results show that the proposed algorithm outperforms its competitors on all or nearly all datasets across multiple performance metrics. We found this consistent pattern of novel models always outperforming competitors particularly noteworthy and worthy of further investigation.

Our research aims to understand this phenomenon by addressing a fundamental question: Is each successive topic model genuinely superior to its predecessors, or is the situation more nuanced than the literature suggests? To answer this, we contribute a comprehensive evaluation framework consisting of ten diverse datasets, eleven topic modelling algorithms, and four performance

metrics. We believe that employing this broader set of datasets, algorithms, and metrics—coupled with the impartial nature of our assessment—will bring a fresh perspective to the realm of CTMs. This approach allows us to examine whether performance advantages are consistent across different contexts or whether they are sensitive to specific data characteristics or evaluation criteria.

2 Topic modelling

This section outlines background information on topic modelling algorithms and topic model performance metrics.

2.1 Overview of topic modelling algorithms

Topic modelling algorithms have evolved significantly over the past three decades, with each advancement enhancing our ability to understand and automatically represent textual content. This section traces the development of topic models with particular focus on contextualised approaches that make use of LLMs.

The foundation of modern topic modelling was established with Latent Semantic Indexing (LSI) [13], which represented documents as vectors and applied singular value decomposition to extract topics. While pioneering, LSI produced topics that lacked interpretability. This limitation was addressed through Non-negative Matrix Factorisation (NMF) [31], which created more interpretable topic-word associations by factorising the document-term matrix into document-topic and topic-word matrices.

A crucial advancement came with probabilistic LSI (pLSI) [22], which reframed topic modelling from a generative perspective. pLSI conceptualised documents as multinomial distributions over topics and topics as multinomial distributions over words. Despite this innovation, pLSI struggled with scalability and could not effectively model new documents outside the training corpus.

These challenges were resolved with Latent Dirichlet Allocation (LDA) [6,7], which introduced a hierarchical structure where document-topic vectors were drawn from a Dirichlet distribution. This approach not only addressed scalability but also enabled modelling of previously unseen documents, making LDA the cornerstone for subsequent developments in the field.

The success of LDA prompted numerous extensions: correlated topic models [5] captured relationships between topics, the Pachinko allocation model integrated word-level correlations [32], multilingual adaptations enabled cross-language analysis [55], and specialised approaches were developed for short documents like social media posts [37]. Another noteworthy development was the introduction of fuzzy clustering techniques [24], which allowed documents partial membership in multiple clusters and demonstrated performance improvements over standard LDA.

A significant paradigm shift occurred with the emergence of neural topic models, which leveraged neural networks to transform complex inference problems

into optimisation tasks. Variational Autoencoders (VAEs) [28] proved instrumental in addressing inference challenges in generative models with latent spaces. Unlike standard autoencoders that compress input into lower-dimensional vectors, VAEs output parameters for probability distributions, enabling more sophisticated data representations.

Building on the VAE framework, the neural variational document model [35] was specifically adapted for topic modelling, using bag-of-words input for the encoder and generating softmax distributions across the vocabulary. Later refinements incorporated various distribution types [34] and adaptations for improved topic coherence [17].

ProdLDA [48] further improved neural topic modelling through three key innovations: eliminating discrete variables that complicate neural network training, employing a variational distribution approximating the Dirichlet distribution, and incorporating dropout layers for improved model robustness and generalisation capabilities.

A fundamental limitation of aforementioned topic models lies in their treatment of words as discrete symbols without semantic understanding. Words like "bank" are treated identically regardless of context (e.g., "river bank" versus "bank robber"). Contextualised topic models address this limitation by leveraging embeddings from LLMs, which capture rich word semantics and contextual nuances.

Language models serve as AI systems specialised in understanding and generating human language [44]. These deep neural networks learn language by training on vast text corpora, encoding semantic knowledge within their weights and connections [15]. This knowledge enables contextually relevant text processing and generation [40,42].

The transformer architecture [52] revolutionised language modelling through self-attention mechanisms, allowing each word to dynamically weigh its importance relative to others in a sentence. This innovation facilitated the development of powerful LLMs, including BERT [15], GPT [42], PaLM [12], and Llama [51], each advancing the state of the art through architectural improvements or enhanced training methodologies.

The knowledge embedded in these models becomes accessible through network activations, particularly as embeddings that condense linguistic entities into numerical vectors [39]. These embeddings relate words based on contextual usage, enabling semantic similarity assessments and sophisticated Natural Language Processing (NLP) tasks.

Contextualised topic models use these LLM embeddings to generate more semantically coherent and contextually relevant topics. Bianchi et al. [4,3] developed ZeroShotTM and CombinedTM by adapting ProdLDA to use LLM embeddings. ZeroShotTM substitutes traditional representations with SBERT embeddings, enabling multilingual topic modelling capabilities. CombinedTM concatenates LLM embeddings with bag-of-words representations, leveraging both contextual understanding and explicit word frequency information.

For short-text analysis, particularly challenging due to limited content, Akash et al. [1] make use of an innovative approach that uses GPT-2 to extend text length before applying topic modelling techniques. Social media analysis has been enhanced by incorporating variational inference into adapted LDA models, generating both topic vectors and user vectors to understand preferences [33].

Sia et al. [46] proposed another method that uses clustered BERT word embeddings to form topic-word associations. Their method was improved by Grootendorst [20] in the BERTopic algorithm. BERTopic makes use of SBERT and establishes document-topic relationships through incorporating dimensionality reduction, to optimise clustering, which allows identification of significant words within clusters.

MPTopic [54] adopts a similar framework to BERTopic but uses MPNet [47] instead of SBERT. The approach of generating embeddings, reducing dimensionality, and applying clustering has been systematically evaluated [18], with findings confirming BERTopic's methodology as particularly effective.

Kardos et al. [25] introduced S^3, a signal processing approach that decomposes semantic embeddings to identify distinct topical signals. S^3 uses independent component analysis to separate semantic signals from document embeddings. They also introduce GMM, a probabilistic clustering approach applied to contextualised embeddings, incorporating optional PCA dimensionality reduction and Dirichlet priors [25]. Kristensen-McLachlan et al. [29] also make use of a decompositon approach in KeyNMF. They use NMF to decompose a document-word similarity matrix.

The evolution from traditional to contextualised topic models demonstrates the field's continuous integration of advances in natural language processing. By leveraging the semantic richness of pre-trained language models, contextualised approaches enable more nuanced analysis of textual content across diverse domains and applications, more closely approximating human understanding of language and meaning.

2.2 Performance metrics

Measuring the performance of topic models is important for assessing their ability to extract meaningful and coherent themes from text data. This evaluation is challenging due to the unsupervised nature of the task. A primary focus is topic coherence, which gauges the semantic interpretability of words within a single topic. Traditional metrics like Normalised Pointwise Mutual Information (NPMI) [8] assess coherence by checking word co-occurrence in external corpora. However, NPMI can be computationally expensive and its alignment with human judgment is debated [23,30]. These limitations have spurred the development of alternative approaches, particularly those using word embeddings.

Word embedding-based coherence metrics, such as Word Embedding-based Centroid Similarity (WECS) [49], creates a weighted vector for each topic based on its word probabilities and embeddings, then measures similarity between these topic vectors; high inter-topic similarity implies low model diversity. Word Embedding-Based Pairwise Similarity (WEPS) [49] also known as Embedding

Coherence (COH) [50], offers a way to capture semantic relationships more directly. WEPS calculates the average pairwise similarity (e.g., cosine similarity) between the vector representations of the top words within a topic. Such metrics can potentially operate in multilingual contexts if appropriate embeddings are used, offering an advantage over traditional co-occurrence measures.

Thielmann et al. [50] propose intruder-based metrics, such as Intruder Shift (ISH) to evaluate coherence and diversity by assessing the impact or identifiability of an "intruder" word introduced into a topic. ISH works by calculating a topic centroid like WECS, then a single word is replaced with a word from another topic known as an intruder word. The centroid is recalculated and the cosine similarity between the original and new centroid is computed. A high similarity indicates that the intruder word did not significantly alter the topic's semantic representation, suggesting that the topic is less diverse or coherent. These newer approaches aim to provide a more nuanced understanding of topic model quality, particularly for models that might generate more abstract or novel thematic representations.

3 Experimental setup

This study presents a comprehensive evaluation of topic modelling approaches across multiple datasets and performance metrics. The experimental framework consists of a systematic pipeline encompassing data ingestion, preprocessing, feature extraction, model training, and evaluation.

3.1 Models

Eight distinct topic modelling algorithms were evaluated in this study, representing different methodological approaches to topic discovery. The implementation details and configurations for each model are listed below.

1. LDA: Used Gensim's [43] multicore implementation with 4 workers
2. NMF: Used scikit-learn's [38] NMF implementation with nonnegative double singular value decomposition initialisation and 200 max iterations
3. ZeroShotTM: Used the Turftopic [25] implementation with a batch size of 32,000 and OpenAI's text-embedding-3-small model for embeddings
4. ZeroShotTM-sbert: Used the same implementation as ZeroShotTM, but used paraphrase-multilingual-MiniLM-L12-v2 embedding model from SBERT [44] instead of OpenAI embeddings
5. CombinedTM: Used the Turftopic [25] implementation with a batch size of 4,000 and OpenAI embeddings
6. CombinedTM-sbert: Used the same implementation as CombinedTM, but with SBERT embeddings
7. BERTopic: Used the Turftopic [25] implementation and the default pipeline configuration with OpenAI embeddings
8. BERTopic-sbert: Used the same implementation as BERTopic, but with SBERT embeddings

9. Gaussian Mixture Model (GMM): Used the Turftopic [25] implementation with the paraphrase-multilingual-MiniLM-L12-v2 embedding model from SBERT [44]
10. KeyNMF: Used the Turftopic [25] implementation with the paraphrase-multilingual-MiniLM-L12-v2 embedding model from SBERT [44]
11. S^3: Used the Turftopic [25] implementation with the paraphrase-multilingual-MiniLM-L12-v2 embedding model from SBERT [44]

Each model was run 10 times per corpus–topic configuration to ensure statistical reliability, with topic numbers ranging from 10 to 200 topics (10, 20, 50, 100, 200). Each dataset–topic number combination is considered to be its own problem instance.

3.2 Datasets

Ten diverse text corpora were employed to evaluate model performance across different domains and document characteristics:

1. 20 Newsgroups: A collection of approximately 20,000 newsgroup posts distributed across 20 discussion categories. Documents were preprocessed by removing headers, footers, and quotes, with a vocabulary of 10,000 most frequent terms and minimum document length of 5 words.
2. IMDB Movie Reviews: A sentiment analysis dataset containing movie reviews with binary sentiment labels. The corpus was processed with stricter filtering criteria, requiring minimum 150 words per document and using 15,000 vocabulary terms to capture domain-specific language patterns.
3. Wikipedia Sample: A curated collection of Wikipedia articles, utilising 20,000 vocabulary terms with minimum 10 words per document to ensure content richness.
4. TREC Questions: A question classification dataset containing questions categorised into semantic types. Due to the short nature of questions, minimal filtering was applied (2 words minimum) with 5,000 vocabulary terms.
5. Twitter Financial News: A specialised corpus of financial news tweets, processed with 8,000 vocabulary terms and minimum 3 words per document to accommodate the concise nature of social media content.
6. PubMed MultiLabel: A dataset comprising scientific paper abstracts from PubMed, used for multi-label text classification.
7. Patent Classification: A dataset consisting of patent abstracts, utilised for patent classification tasks.
8. Goodreads Book Genres: A corpus of book descriptions from Goodreads, used for genre classification.
9. Battery Abstracts: A collection of research paper abstracts related to battery data, focusing on scientific text analysis.
10. T2-RAGBench ConvFinQA: A conversational financial question answering dataset, designed for evaluating retrieval-augmented generation models in a financial context.

All corpora underwent systematic preprocessing including tokenisation, lowercasing, stopword removal, lemmatisation, and vocabulary filtering based on document frequency thresholds tailored to each dataset's characteristics. Document chunking was performed with a maximum token limit of 8,190 and 128 to accommodate model input constraints for OpenAI and SBERT, respectively.

3.3 Performance metrics

Four evaluation metrics were employed to assess topic quality from different perspectives: NPMI, WEPS, WECS and ISH. All metrics were computed for the top 10 words per topic, and evaluations were performed systematically across all model-corpus combinations. We report the negative ISH (NISH) to facilitate comparison with other metrics, as this allows all metrics to share the same maximum direction (i.e. higher is better).

4 Results and discussion

Running the eleven algorithms on the ten datasets for each of the five topic numbers for ten repeats resulted in a total of 5,500 model runs. This is significantly more than other experiments in literature [3,4,6,7,20,25]. The runs were then evaluated using the four performance metrics, resulting in a total of 22,000 metric evaluations. This section presents different views of the results tensor: 11 algorithms $\times$ 50 problem instances $\times$ 4 metrics $\times$ 10 independant runs, where 50 problem instances = 10 datasets $\times$ 5 number of topics. Since the results are of too high dimensionality to be presented in a single table, we present the results through four complementary analyses that collectively reveal the nuanced nature of algorithm performance.

Table 1 presents the overall performance of each algorithm across all problem instances and metrics. While this summary view provides a convenient ranking of algorithms, it fundamentally obscures the nuanced performance patterns that emerge when examining individual problem instances. The table shows that different algorithms excel at different metrics. However, this aggregate view masks the reality that algorithm performance varies significantly across different datasets and topic numbers, losing critical information about when and why specific algorithms perform better.

Figure 1 reveals a more nuanced picture through pairwise comparisons, showing how many times each algorithm significantly outperforms another across all 50 problem instances. These heatmaps demonstrate that algorithmic dominance is not uniform, and that different algorithms excel in different contexts. This heterogeneous performance landscape suggests that no single algorithm consistently dominates across all evaluation scenarios, indicating that the choice of "best" algorithm depends heavily on the specific problem context and performance criteria prioritised by the user.

When dealing with multiple performance metrics simultaneously, the concept of optimality becomes more complex. In multi-objective optimisation, *Pareto*

Table 1. Overall performance summary for each performance metric. **Bold** indicates statistically significant best performer(s), underlined indicates statistically significant second-best performer(s) (Mann-Whitney U test, $p < 0.05$).

	NPMI	WEPS	WECS	NISH
LDA	-0.047 ± 0.255	-0.119 ± 0.026	-0.121 ± 0.024	0.146 ± 0.028
NMF	**0.136 ± 0.191**	-0.099 ± 0.018	-0.122 ± 0.023	0.145 ± 0.026
ZeroShotTM	-0.011 ± 0.310	-0.093 ± 0.012	**-0.105 ± 0.013**	0.127 ± 0.016
ZeroShotTM-sbert	0.018 ± 0.302	-0.094 ± 0.011	-0.109 ± 0.016	0.131 ± 0.017
CombinedTM	0.019 ± 0.290	-0.097 ± 0.015	-0.111 ± 0.017	0.134 ± 0.020
CombinedTM-sbert	0.037 ± 0.274	-0.097 ± 0.015	-0.113 ± 0.018	0.136 ± 0.021
BERTopic	0.080 ± 0.223	-0.100 ± 0.015	-0.118 ± 0.019	0.141 ± 0.021
BERTopic-sbert	0.075 ± 0.224	-0.100 ± 0.014	-0.119 ± 0.018	0.142 ± 0.021
GMM	0.107 ± 0.242	-0.103 ± 0.017	-0.129 ± 0.024	**0.152 ± 0.027**
KeyNMF	0.083 ± 0.243	-0.088 ± 0.012	-0.116 ± 0.021	0.132 ± 0.021
S^3	-0.173 ± 0.347	**-0.084 ± 0.012**	-0.113 ± 0.015	0.130 ± 0.017

dominance provides a principled framework for comparing solutions. Solution A Pareto dominates Solution B if A is at least as good as B in all objectives and strictly better in at least one objective. For example, say we are evaluating topic models using three metrics, where higher is better for each metric. For Solution A to Pareto dominate Solution B, two conditions must be satisfied:

1. Solution A must perform at least as well as Solution B across every single metric.
2. Solution A must perform strictly better than Solution B in at least one metric.

If Solution A performs worse than Solution B in even a single objective, then A cannot Pareto dominate B, regardless of how much better it might perform in other objectives. The set of all non-dominated solutions forms the *Pareto frontier*, representing the optimal trade-offs between the competing objectives. In the context of topic modelling, this means that a model is Pareto optimal if no other model simultaneously achieves better performance across all metrics.

Figure 2 illustrates this concept by showing Pareto frontiers for different metric pairs across all problem instances. The plots reveal that multiple algorithms simultaneously occupy the Pareto frontier, with each offering distinct trade-offs between objectives. This visualisation demonstrates that once we move beyond single-metric evaluation, the notion of a single "best" algorithm becomes meaningless. Instead, we must consider which trade-offs align with our priorities. This exemplifies how different algorithms can be simultaneously optimal depending on the relative importance a user places on different performance aspects.

Table 2 presents a systematic analysis of Pareto optimality across all 50 problem instances using all four metrics simultaneously. Remarkably, the results show that in 42 out of 50 problem instances (84%), all algorithms are Pareto optimal. This striking finding indicates that across the vast majority of evaluation scenarios, every algorithm offers a unique combination of strengths and weaknesses

that cannot be dominated by any other algorithm. Even in the 8 instances where some algorithms are dominated, the number of dominated algorithms is minimal (1-3 out of 11), suggesting that most algorithms contribute meaningfully to the Pareto frontier.

Table 2. Number of non-Pareto optimal models for each dataset-topic combination. Pareto optimality calculated using all four metrics (NPMI, WEPS, WECS, NISH).

Dataset	Number of topics				
	10	20	50	100	200
battery-abstracts	1	1	0	2	0
goodreads-bookgenres	0	0	0	0	0
imdb-reviews	0	0	0	0	0
newsgroups	0	0	0	2	0
patent-classification	0	0	0	0	0
pubmed-multilabel	0	0	0	0	0
t2-ragbench-convfinqa	1	1	0	1	3
trec-questions	0	0	0	0	0
twitter-financial-news	1	0	1	0	0
wikipedia-sample	0	0	0	0	0

This comprehensive analysis provides compelling evidence that the topic modelling performance landscape is characterised by fundamental trade-offs rather than clear hierarchies. The pervasive presence of multiple algorithms on the Pareto frontier demonstrates that different algorithms excel under different conditions and evaluation criteria, making the question "which algorithm is best?" fundamentally unanswerable without additional context about user priorities and problem characteristics.

The results presented above collectively demonstrate what is known in the algorithmic literature as performance complementarity [26]. This phenomenon, observed across numerous computational domains, describes the situation where different algorithms perform optimally on different types of problem instances, with no single algorithm dominating across all scenarios. As noted by Kerschke et al. [26], this complementarity has been observed for practically all NP-hard decision and optimisation problems, including propositional satisfiability, constraint satisfaction, planning and scheduling problems, and many others.

Our evaluation reveals that contextual topic modelling exhibits clear performance complementarity. The evidence supporting this conclusion includes: (1) the aggregate performance table showing different algorithms excelling at different metrics, (2) the pairwise dominance heatmaps revealing heterogeneous winning patterns across problem instances, (3) the Pareto frontier analysis demonstrating that multiple algorithms simultaneously represent optimal trade-offs, and (4) the comprehensive Pareto analysis showing that most algorithms remain non-dominated even when considering all metrics simultaneously.

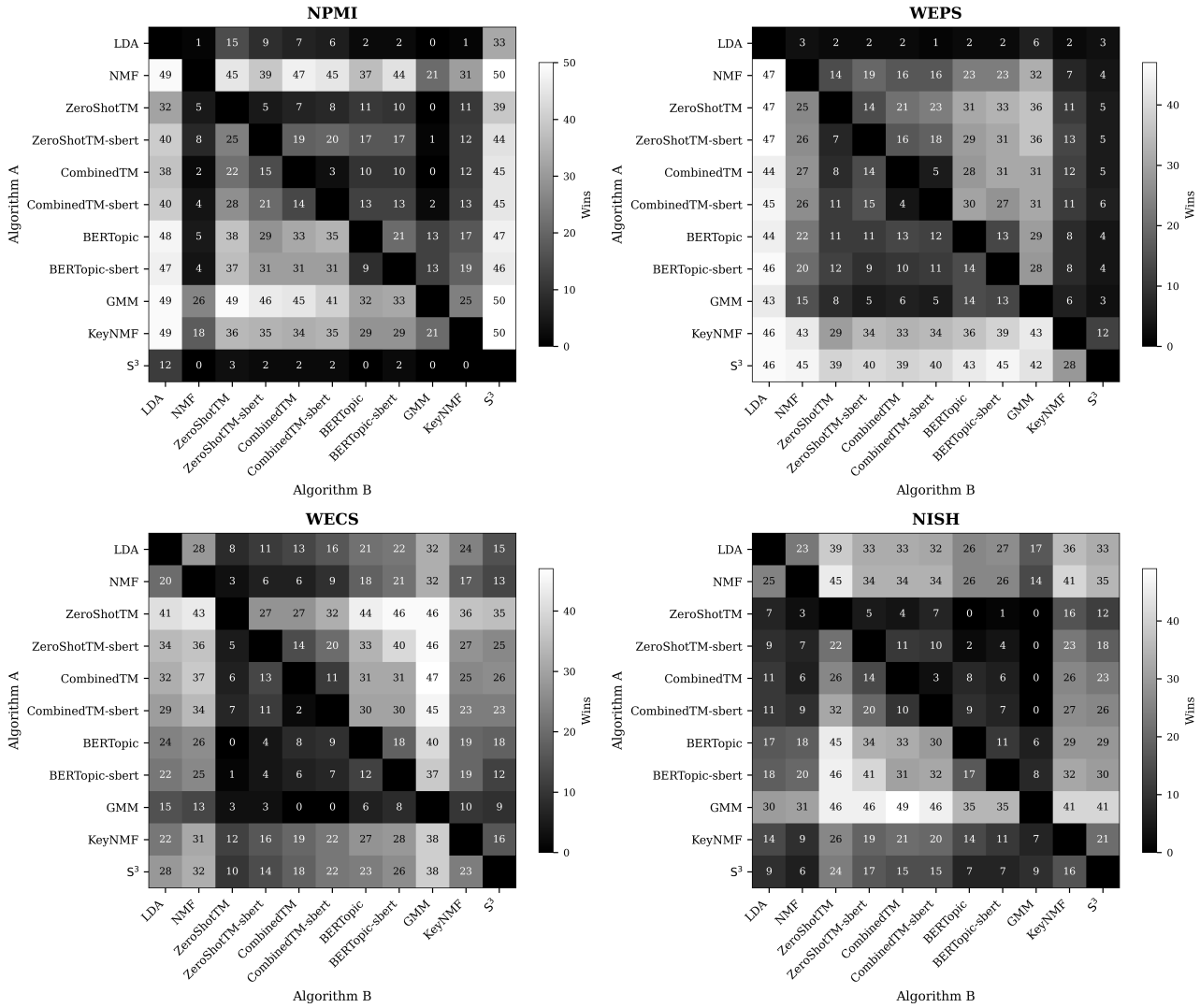

Fig. 1. Algorithm comparison heatmaps showing the number of statistically significant wins for each algorithm pair across all metrics. Each cell (i,j) shows how many times Algorithm A (row) significantly outperformed Algorithm B (column) across all 50 dataset-topic combinations. Statistical significance determined using Mann-Whitney U test ($p < 0.05$). Lighter shades indicate more wins.

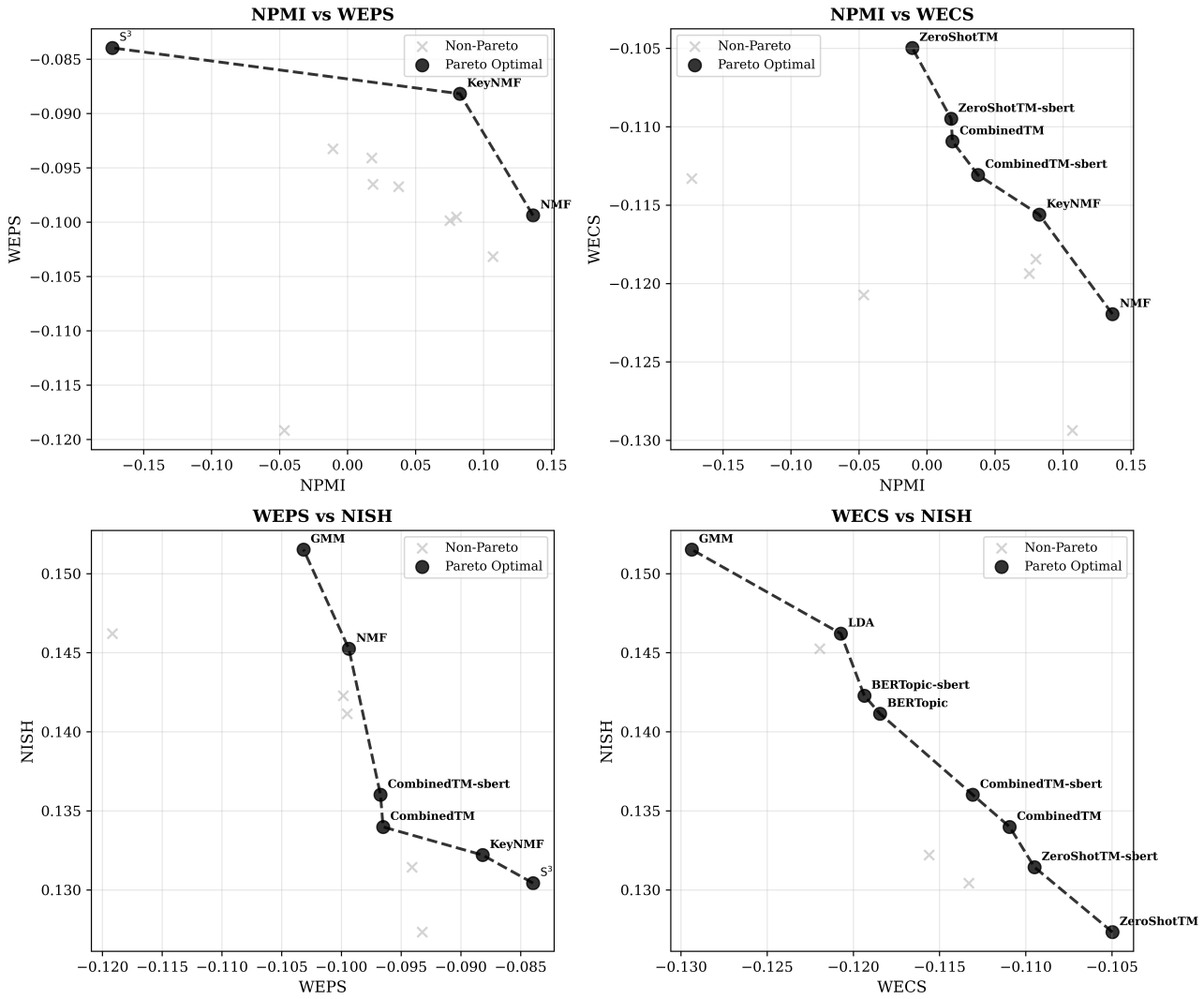

Fig. 2. Pareto frontier analysis for metric pairs showing optimal trade-offs between different evaluation criteria. Black circles connected by dashed lines represent Pareto optimal models that are not dominated by any other model across both metrics. Gray crosses show non-Pareto optimal models. Top-left: NPMI vs WEPS coherence measures. Top-right: NPMI vs WECS coherence measures. Bottom-left: WEPS vs NISH (topic diversity). Bottom-right: WECS vs NISH (topic diversity).

5 Conclusion

This study addresses the fundamental question of which contextual topic modelling algorithm is best through a comprehensive empirical evaluation of eleven algorithms across ten datasets, five numbers of topics, and four performance metrics. Our findings reveal that the answer to this question is more nuanced than the literature suggests.

Rather than identifying a single superior algorithm, our results demonstrate clear evidence of performance complementarity in contextual topic modelling. The most striking finding is that in 84% of problem instances, all algorithms are Pareto optimal when considering all four metrics simultaneously, indicating that each algorithm offers a unique combination of strengths that cannot be dominated by others. The results also indicate that the field should focus on understanding when and why different algorithms perform well, potentially leading to algorithm selection strategies that match algorithms to problem characteristics. Furthermore, these findings suggest that the common practice in literature of proposing new algorithms that "outperform" existing ones may be misleading, since what appears to be superiority may simply reflect performance complementarity across different evaluation scenarios. A more honest and comprehensive evaluation, as presented here, reveals the existence of algorithmic performance in topic modelling.

The practical implications are equally important. For practitioners, our findings indicate that algorithm selection should be guided by specific problem characteristics and performance priorities rather than blanket recommendations. Different algorithms offer different trade-offs between coherence measures (NPMI, WEPS, WECS) and diversity (ISH), and the optimal choice depends on the relative importance a user places on these different aspects of topic quality.

Our study contributes to the growing recognition that performance complementarity is a fundamental characteristic of computational problems, extending this understanding to the domain of contextual topic modelling. This work provides a foundation for future research into algorithm selection strategies that can automatically match algorithms to problem characteristics, potentially leading to more robust and reliable topic modelling in practice.

Future research should explore the development of algorithm selection frameworks that can predict which algorithm will perform best for a given dataset and evaluation criterion combination. Additionally, investigating the underlying problem characteristics that drive performance complementarity could provide deeper insights into when and why different algorithms excel. Finally, the comprehensive evaluation framework developed in this study can serve as a benchmark for future topic modelling research, encouraging more honest and complete performance assessments.

Acknowledgements. The authors would like acknowledge the contributions of George Davie, as well as the funding from Isazi Consulting for the student's PhD.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akash, P.S., Das, T., Chang, K.C.C.: TopicAdapt- An Inter-Corpora Topics Adaptation Approach (arXiv:2310.04978) (Oct 2023)
2. Akash, P.S., Huang, J., Chang, K.C.C.: Let the Pretrained Language Models "Imagine" for Short Texts Topic Modeling (arXiv:2310.15420) (Oct 2023)
3. Bianchi, F., Terragni, S., Hovy, D.: Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (Jun 2021)
4. Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E.: Cross-lingual Contextualized Topic Models with Zero-shot Learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1676–1683. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.eacl-main.143
5. Blei, D., Lafferty, J.: Correlated topic models. Advances in Neural Information Processing Systems **18**, 147 (2006)
6. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Advances in Neural Information Processing Systems **14** (2001)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**(Jan), 993–1022 (2003)
8. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of the Biennial GSCL Conference. pp. 31–40 (2009)
9. Cao, H.: Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark (arXiv:2406.01607) (Jun 2024)
10. Chen, X., He, T., Hu, X., Zhou, Y., An, Y., Wu, X.: Estimating Functional Groups in Human Gut Microbiome With Probabilistic Topic Models. IEEE Transactions on NanoBioscience **11**(3), 203–215 (Sep 2012). https://doi.org/10.1109/TNB.2012.2212204
11. Chen, X., Hu, X., Lim, T.Y., Shen, X., Park, E.K., Rosen, G.L.: Exploiting the Functional and Taxonomic Structure of Genomic Data by Probabilistic Topic Modeling. IEEE/ACM Transactions on Computational Biology and Bioinformatics **9**(4), 980–991 (Jul 2012). https://doi.org/10.1109/TCBB.2011.113
12. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
13. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Beck, L.: Improving information-retrieval with latent semantic indexing. In: Proceedings of the ASIS Annual Meeting. vol. 25, pp. 36–40. Information today inc 143 (1988)
14. Denti, T., Sun, W., Ji, S., Moen, H., Kerro, O., Rannikko, A., Marttinen, P., Koskinen, M.: Using unsupervised topic modeling to uncover document hierarchy and latent topics in prostate cancer clinical texts. Preprint, MedRxiv (Jan 2024). https://doi.org/10.1101/2024.01.29.24301349
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)

16. Dieu, D.T., Dinh, D.: Using Topic in Summarization for Vietnamese Paragraph. International Journal of Advanced Computer Science and Applications **14**(10) (2023). https://doi.org/10.14569/IJACSA.2023.0141078
17. Ding, R., Nallapati, R., Xiang, B.: Coherence-Aware Neural Topic Modeling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 830–836. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1096
18. Eklund, A., Forsman, M., Drewes, F.: An Empirical Configuration Study of a Common Document Clustering Pipeline. Northern European Journal of Language Technology **9** (2023)
19. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT Sentence Embedding (arXiv:2007.01852) (Mar 2022)
20. Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arXiv:2203.05794) (Mar 2022)
21. Gurusamy, B.M., Rengarajan, P.K., Srinivasan, P.: A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer. International Journal of Electrical and Computer Engineering (IJECE) **13**(6), 6663 (Dec 2023). https://doi.org/10.11591/ijece.v13i6.pp6663-6672
22. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 50–57 (1999)
23. Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., Resnik, P.: Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. In: Proceedings of Neural Information Processing System. pp. 2018–2033 (2021)
24. Karami, A., Gangopadhyay, A., Zhou, B., Kharrazi, H.: Fuzzy Approach Topic Discovery in Health and Medical Corpora. International Journal of Fuzzy Systems **20**(4), 1334–1345 (Apr 2018). https://doi.org/10.1007/s40815-017-0327-9
25. Kardos, M., Kostkan, J., Vermillet, A.Q., Nielbo, K., Enevoldsen, K., Rocca, R.: Semantic Signal Separation (arXiv:2406.09556) (Jun 2024)
26. Kerschke, P., Hoos, H.H., Neumann, F., Trautmann, H.: Automated Algorithm Selection: Survey and Perspectives. Evolutionary Computation **27**(1), 3–45 (Mar 2019). https://doi.org/10.1162/evco_a_00242
27. Khaled, E., Omar, Y.M.K., Hodhod, R.: Towards an Enhanced Model For Contextual Topic Identification. Proceedings of NILES2023: 5th Novel Intelligent and Leading Emerging Sciences Conference pp. 188–193 (2023)
28. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: The International Conference on Learning Representations (ICLR). Banff (2014). https://doi.org/10.48550/arXiv.1312.6114
29. Kristensen-McLachlan, R.D., Hicke, R.M.M., Kardos, M., Thunø, M.: Context is Key(NMF): Modelling Topical Information Dynamics in Chinese Diaspora Media (arXiv:2410.12791) (Oct 2024). https://doi.org/10.48550/arXiv.2410.12791
30. Lau, J.H., Newman, D., Baldwin, T.: Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 530–539. Association for Computational Linguistics, Gothenburg, Sweden (2014). https://doi.org/10.3115/v1/E14-1056
31. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (Oct 1999). https://doi.org/10.1038/44565

32. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 577–584 (2006)
33. Liu, L., Lin, Q., Tong, H., Zhu, H., Liu, K., Wang, M., Zhang, C.: Neural Personalized Topic Modeling for Mining User Preferences on Social Media. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 1545–1555. ACM, Birmingham United Kingdom (Oct 2023). https://doi.org/10.1145/3583780.3614987
34. Miao, Y., Grefenstette, E., Blunsom, P.: Discovering Discrete Latent Topics with Neural Variational Inference. International Conference on Machine Learning pp. 2410–2419 (2017)
35. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: International Conference on Machine Learning. pp. 1727–1736. PMLR (2016)
36. Mim, S.S., Logofatu, D., Guerrero-Contreras, G., Medina-Bulo, I.: Leveraging Topic Modeling and Extractive Summarization for Unlocking Insights from NeurIPS Papers. In: 2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). pp. 1–6. IEEE (2024)
37. Paul, M.J., Dredze, M.: Discovering Health Topics in Social Media Using Topic Models. PLoS ONE **9**(8), e103408 (Aug 2014). https://doi.org/10.1371/journal.pone.0103408
38. Pedregosa, F., Pedregosa, F., Varoquaux, G., Varoquaux, G., Org, N., Gramfort, A., Gramfort, A., Michel, V., Michel, V., Fr, L., Thirion, B., Thirion, B., Grisel, O., Grisel, O., Blondel, M., Prettenhofer, P., Prettenhofer, P., Weiss, R., Dubourg, V., Dubourg, V., Vanderplas, J., Passos, A., Tp, A., Cournapeau, D.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12** (2011)
39. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations (arXiv:1802.05365) (Mar 2018)
40. Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language Models as Knowledge Bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2463–2473. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1250
41. Pylov, P., Maitak, R., Protodyakonov, A.: The Latent Dirichlet Allocation (LDA) generative model for automating process of rendering judicial decisions. E3S Web of Conferences **431**, 05005 (2023). https://doi.org/10.1051/e3sconf/202343105005
42. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training. https://openai.com/research/language-unsupervised (2018)
43. Řehůřek, R., Sojka, P.: Gensim-python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**(2) (2011)
44. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Conference on Empirical Methods in Natural Language Processing (2019)
45. Rogers, A., Kovaleva, O., Rumshisky, A.: A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics **8**, 842–866 (Dec 2020). https://doi.org/10.1162/tacl_a_00349

46. Sia, S., Dalmia, A., Mielke, S.J.: Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1728–1736. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.135
47. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems **33**, 16857–16867 (2020)
48. Srivastava, A., Sutton, C.: Autoencoding Variational Inference For Topic Models. In: 5th International Conference on Learning Representations. Toulon, France (Mar 2017)
49. Terragni, S., Fersini, E., Messina, E.: Word Embedding-Based Topic Similarity Measures. In: Métais, E., Meziane, F., Horacek, H., Kapetanios, E. (eds.) Natural Language Processing and Information Systems, vol. 12801, pp. 33–45. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-80599-9_4
50. Thielmann, A., Reuter, A., Seifert, Q., Bergherr, E., Säfken, B.: Topics in the Haystack: Enhancing Topic Quality through Corpus Expansion. Computational Linguistics pp. 1–36 (Jan 2024). https://doi.org/10.1162/coli_a_00506
51. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971) (Feb 2023)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. Conference and Workshop on Neural Information Processing Systems (2017)
53. Yoo, Y., Choi, J.: Topic-VQ-VAE: Leveraging latent codebooks for flexible topic-guided document generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19422–19430 (2024)
54. Zhang, X., Milios, E.: MPTopic: Improving topic modeling via Masked Permuted pre-training (arXiv:2309.01015) (Sep 2023)
55. Zhao, B., Xing, E.: HM-BiTAM: Bilingual topic exploration, word alignment, and translation. Advances in Neural Information Processing Systems **20** (2007)