Testing the effects of parameter manipulation on BERTopic

Thomas Compton
University of York
York, United Kingdom
thomas.compton@york.ac.uk

Abstract

Topic models are drawing increased commercial and academic interest for their ability to summarise large quantities of unstructured data (Hosseiny Marani and Baumer, 2023; Egger and Yu, 2022; Li et al., 2023; Madrid-García et al., 2024). As unsupervised machine learning approaches, they offer a method of studying data for researchers and understanding the important points of large amounts of text to general users. At the same time, they risk becoming a 'black box' (Wang et al., 2024), where researchers add their data and uncritically assume the output results are a 'correct' summarisation of the data. In this article, I wish to approach the evaluation of topic models from a database perspective drawing inferences from outputs from 1140 runs of BERTopic models. The goal will be to explore potential trade-offs in optimising certain aspects of the topic model and to consider what these findings mean for how we should interpret topic models.

1 Introduction

First, is it important to clarify the architecture of topic models. With BERTopic, for example, the approach is stochastic (Kumar, Karamchandani and Singh, 2024), relying on UMAP and HDBScan (Gokcimen and Das, 2024; Grootendorst, 2022). UMAP is used to reduce the embeddings, which can be produced by sentence transformers. Then HDBScan clusters the vectors into meaningfully distinct groups (Lin et al., 2024). For research, I shall focus on UMAP and HDBScan because they are the sources of variation in the runs. Whereas, the usage of Count Vectorisation to extract ngrams is deterministic, meaning that it will produce the same top 10 ngrams for the same cluster.

It is important to clarify BERTopic's advantage against a deterministic model such as K means. If one were to use K means clustering, the variation would come from the user inputting the K value which is the number of clusters preset. After this has been set, it should be the case that rerunning the model produces the same results. However, the trade-off is that K means do not necessarily provide meaningful topics. So, HDBScan is sacrificing repeatability for better quality clusters. The question emerges from this, how much of an issue is repeatability and how improved is HDBScan from K means.

BERTopic has the advantage of being able to be controlled by users through the parameters (Liu, 2024). Table 1 could illustrate the relationship between min_cluster_size which is a parameter in HDBScan used to fine tune the approach to have larger clusters. Min_topic_size is used in the BERTopic wrapper to manually set a limit for the minimum number of sentences in a topic. N_neighbo[u]rs is used in UMAP to set the nearest neighbours. If these were deterministic approaches, then this table would be a useful indication of the best parameters. Yet, because it is not deterministic, re-running the model with the same parameters will produce different results. Therefore, this table sorted by error_size which is the number of sentences categorised into the -1 topic by BERTopic meaning that the model was unable to find a topic for them.

2 Existing Studies On Topic Model Evaluation

Given the popularity of topic models and range of potential options it is unsurprising to find many comparisons between topic models such as BERtopic, LDA, and/or Top2Vec (An, Oh and Lee, 2023; Egger and Yu, 2022; Li et al., 2025).

Table 2 suggests that BERTopic users are using different methods to evaluate their model with no clear most prominent metric in this sample. PUV (Pairwise Uniqueness Value) did not appear in the corpus. This metric evaluates the overlap in ngrams between topics. If topics share many ngrams, it suggests each topic is not distinct and therefore the clustering was not effective in semantic differentiation. PMI, which can be normalised, compares the probabilities of the ngrams co-occuring in the corpus against each ngrams probability of occurring in the corpus (Ranaweera et al., 2025). Therefore, this is not comparing topics, but an attempt to guage how effectively the topic model has grouped together co-occuring terms. However, BERTopic uses C-TF-IDF, which means that it is not outputting the most common ngrams from each cluster by attempting to find contextually significant terms. This means that this co-occurance may suffer as BERTopic may be choosing low

frequency terms as an internal buffer against topic overlap, especially in smaller corpora.

Gan et al., (2024) uses cosine similarity to compare the semantic differences between topics (WE). This can be done through converting the ngrams output for each topic into a string then embedding it with sentence transformers. After this the distance between the vectors of each pseudo-sentence can be compared. Alternatively, exach ngram could be converted into a vector and the usual PUV approach can be used with the advantage of increased sensitivity to semantic similarity. This method is computationally intensive, as compared to PUV. The issue is that each model uses a different approach to extract ngrams, so the issue becomes are we testing these packages as wrappers or should the focus be normalising the ngram selection process to ensure that the clustering process is being tested.

Equally, topic coherence values can be computed using a similar approach but comparing ngrams within topics using cosine similarity (Khodeir and Elghannam, 2025). As with PMI and PUV, the focus is on the ngrams which is the output from the models many users may highlight. At the same time, using UMAP projections, users are able to inspect the topic overlaps through looking at the clusters. This approach would allow for inspection of how the effective the clustering process has been, instead of focusing on evaluating the ngram selection. This is valuable because the ngram outputs provide useful information about the topics, but for researchers interested in understanding limitations to the clustering approach, they do not provide much useful information about how effectively the models have been clustering. In this corpora, there was no mention of using Gini Coefficients, for example, to test the distributions of clusters.

Looking at terms associated with topic modelling, it is clear the corpora contains many articles comparing BERTopic to LDA or top2vec. These comparisons could be tests to see which performs better, or they could be triangulations where researchers attempt to gain more information about their corpora by using both. The use of comparative approaches is not surprising as the evaluative metrics seem to focus on testing the outputs of models so can conveniently move between models. Although, this will be limited by how each model uses a different approach at extracting ngrams. Moreover, this approach is a good step because there are no ground truth topic labels, so comparing outputs can increase confidence in findings. On the other hand, if each model output is not optimised, the triangulation may be giving users a false sense of confidence in their findings.

To explain this tension further, I shall look at how BERTopic specifically is being used. While UMAP and HDBScan are being identified as the steps for topic modelling, there is less discussion of the key metrics for fine-tuning the outputs. As we demonstrated in the last section, the changing of these metrics can have a significant effect on the error size. This means that these topic models are theoretically 'plug and play', but in practice, outputs from one dataset could vary in quality. However, to test this, it will be important to establish by what standard we are measuring the denigration of quality. This article will not assume 'plug and play' usage is inherently poor, instead seeking to understand the potential limitations of both manually interference and allowing the model to run without customisation.

3 Developing Evaluation Metrics

Before discussing the metrics I propose for evaluating topic models, it will be important to establish the features of an ideal topic model. On one hand, this is a difficult task because text corpora are non-normalised and highly varied data. Therefore, the expectations cannot be too fixed and risk losing sensitivity to variation. On the other hand, there are metrics based on the model outputs that can resolve this issue. Since the model is already normalising the data and providing a quantitative output, it is possible to focus on the outputs of the model and make comparisons between different models.

The Gini coefficient is widely used to measure inequality in distribution, here applied to sentence-topic assignments. A higher Gini value means more imbalance, indicating that some topics dominate while others are underrepresented

Table 4 explores BERTopic and Top2Vec using 'all-MiniLM-L6-v2', the model which is the default embedding model for bertopic (Gokcimen and Das, 2024).[1] It is referred to in 17% of the literature corpus. Therefore, the goal of this table is to give a baseline of what a user might receive in quality from each model, comparing the top 20 topics from each. To standardise this, the LDA, using Gensim's model, had 20 topics selected. This model compares the top 20 topics from each model, with % appearance referring to the coverage of the model as a percentage of the total sentences. BERtopic scores best in PUV and NUV, demonstrating the model produces distinct ngrams between topics(PUV) and distinct ngrams overall. LDA performs best at topic 20 Size, & Appearance, Coherence and KFS. This means that LDA's topic clusters are larger, which is unsurprising given the high coverage. Yet, this is not providing a trade off in coherence. Although the much increased KFS could indicate the ngram outputs are being dilluted because of the cluster size. Top2Vec has the best Gini, which indicates the the most even distribution of topics, but each value is notably low.

From these metrics, it might be fair to argue BERTopic, tentatively, outperforms Top2Vec by providing more coherent but smaller clusters. On the other hand, these prelimi-

[1]For some users, confusion will arise that this model does not use a BERT model by default. The model 'all-MiniLM-L6-v2', may improve on BERT, but is not BERT (Yin and Zhang, 2024). This means that BERTopic is not using BERT models by default, suggesting the name may draw confusion.

nary metrics do not provide a broader picture of the scope and potential damage this trade-off may cause. What has made BERTopic the model for this study, over Top2Vec, also involves the modularity of the wrapper, as it is designed to facilitated. Compared to Grootendorst (2021), each of these coherence score underperforms his tests. Although, his test of BERTopic uses 'all-mpnet-base-v2' a model with benchmarks superior to 'all-miniLM-L6-v2'. Medvecki et al. (2024, p.6) Found -.058 using the similar 'paraphrase-multilingual-mpnet-base-v2', demonstrating a better score than -0.075 for our test.

These proposed metrics will be explored in the next section to see how they interact with each other and how far they are useful in creating more valuable topic models. As already demonstrated in the previous section, repeating topic models using randomised parameters can lead to optimised metrics at the expense of user time. To ensure continuity, a similar approach will be taken, this time using a larger corpus and 211 runs.

4 Descriptives

The larger corpus has 44 unique values from 301 models. From these metrics, it is clear Gini Coefficient, the size of the 20th topic and error size saw more variation than keyword frequency. Gini and topic 20 size should have a relationship, which will be explored in the next section. Keyword frequencies suggests most models will have reasonable consistency on this metric. However, the Gini coefficient indicates there can be issues with the sentence distribution. Therefore, these data suggest there can be variation between models a user may want to avoid. At the same time, the lack of unique values from this number of runs indicates the perameter range was able to keep the output reasonable consistent.

Therefore, there will be diminishing returns for users that run their more models. The issue for research is that the parameters interact such that even an approach that only used one perameter combination once would still see repeated values.

With the smaller corpus, it was possible to run more models and test a wider range of parameters. This is a potential limitation, as the graph shows, with this approach being unlikely to be chosen by a user and contained a collection of models with less sentence coverage than 40%. Equally, the models generally performed worse than the larger corpus. Therefore, the size of this corpus is likely contributing to the issues.

On the other hand, this dataset is useful because it contains worse models and therefore if there is no trade-off found with this dataset, it would indicate a stronger chance for users to be able to achieve successful models randomly. Therefore, this dataset is a test case of intentionally bad models, whereas the first is a test case of better models. This allows for an approach based on stress testing BERTopic and one more focused on emulating how users may reasonably interact with BERTopic.

5 Bertopic Statistics

Before discussing the findings, it is worth discussing the context of this data. Knowing that BERTopic is a stochastic because of UMAP and HDBScan (Kumar, Karamchandani and Singh, 2024), these tests will be limited by being a partial collection of results. It may be more appropriate to say, it is difficult to gain a full set of possible outputs because of the stochastic nature and variation between datasets. To resolve this issue, I have used 2 datasets of different sizes. This will provide the opportunity to discuss differences between them and possible reasons. The other limitation in this data is the nature of textual corpora. Each corpora is substantially different, especially regarding how far high frequency terms are used within thematically consistent sentences. Essentially, one might expect some corpora to score higher on ngram values because of the content of the sentences, not the BERTopic model. Grootendorst (2022) also found varying outputs form BERTopic between corpora.This means between corpora comparisons should be done with caution, and the focus should be on metrics which should remain reasonably consistent across corpora. Therefore, there are limitations to this statistical approach. However, it should be useful to give guidance to users of BERTopic surrounding general issues, which should not be interpreted as inherent flaws with BERTopic.

Beginning with the Gini Coefficients, this metric is simple to compare across corpora because it is evaluating the model outputs. The small corpora found Spearman's Rho -0.55, $p = 0.0000$. The large corpora found Rho 0.57, $p = 0.0001$. Both results demonstrate a statistically significant relationship between these metrics, indicating that with more sentences in the -1 topic greater imbalances within the number of sentences in each topic. It is reasonable to argue, therefore, that models which have had their coverage of sentences increases by changing the parameters lead to more sentences falling into the top topics at the expense of the latter topics.

The metrics indicate a clear trade-off when maximising the model coverage for gini coefficients. On the other 2 metrics, the relationships was not as meaningful, with there being less reason to argue model coverage would lead to a difference in model quality. Therefore, a user could argue they are willing to sacrifice the a more even distribution of sentences for larger coverage. This argument would be reasonable based on the evidence and the goals of their research. Especially, if a user is interesting in using a UMAP projection, this will be an issue for their approach as the skewness could make this representation less readable.

6 Conclusion

There appears to be variation along the different metrics explored between better and worse models in certain

areas. Equally, the choice of deciding which metrics to optimise lead to trade-offs that might ask greater questions of what topic models are for. Whether topic models exist to explain the most possible data or whether they exist to create the most coherent topic appear to be contradictory aims within this model. It will depend on one's interpretation of explanatory power within a topic model. This is an issue because the number of sentences within the -1 may not be a useful metric. This is because topics above the count of 20 may be excluded by the user or may be of poor quality. Simplistically, the question is whether we are evaluating the topic model as a whole of the top 20 topics. If we are evaluating with the use cases in mind, it would seem strange to be evaluating beyond the 20 mark. This is why this article has focused on evaluations of the top 20. Equally, because of how the sentence counts can decrease already by the 20th topic, these top 20 are liable to include the majority of topics. With this in context, the damage of not adding topics past the 20th to the -1 number of sentences may not be such an issue. On the other hand, the issue is not entirely quantitative. It could be possible that the -1 is a reflection that every corpora will have low value sentences. The issue with this argument is that it would be impossible to create a standard for this. Semantic outliers can easily be identified but these are not necessarily low value. What we do not know from the -1 category is how appropriate that count is. It is feasible that it could be that each corpora should have a number of sentences in this category demonstrating no overfitting.

7 Limitations

If part of the gambit of this article was exploring whether it is possible to statistically evaluate BERTopic models, it would be fair to state the results are mixed. The relationship between the number of sentences in the -1 topic and gini coefficient was clear. An issue is that repeated runs produce similar results. Where this approach suffers is that the error size essentially serves as an index for each model with the error size dictating the results of the metrics consistently. Essentially, a model that has the same error size is the same model. This is because parameters do not dictate the error size. Therefore, the only way to find the error size is the run the model. So, it would be impossible to pre-establish the runtime for an approach of this sort and to approach an exhaustive list of possible outputs.

Moreover, there is a danger of creating a dataset of unusable models to test BERTopic. This would lead to an analysis on models users would most likely reject. Therefore, it might be evaluating BERTopic as a package, but not evaluating possible models produced by users. That is, we would be creating models for the sake of a dataset instead of considering what sort of variation a user might experience in their usage of BERTopic. This is the advantage of the approach taken in this article. We have used possible models that a user might reasonably consider for their project. We have not analysed models a user might reject, with our approach not storing outputs from models with less than 20 topics

The evaluation of BERTopic models cannot lose sight of the purpose of BERTopic models. Moreover, these metrics are designed to provide insights but to not dictate what a 'good' model is. This decision will be, ultimately, qualitative. Furthermore, any attempt to evaluate BERTopic models will be time consuming and inefficient. For most users, an approach which includes 5 runs and chooses the best may be what reduces their inconvenience. At the same time, it is plausible 5 runs could include similar values. Ultimately, this is the inherent flaw with using a stochastic model (Kumar, Karamchandani and Singh, 2024), that the chance will always be that the comparatively best model will not be the 'best' model because it is always impossible or at least difficult to justify the runtime to find the 'best' in a large pool of possible results.

References

An, Y., Oh, H., & Lee, J. (2023). Marketing insights from reviews using topic modeling with bertopic and deep clustering network [[Online]. Accessed: 18 July 2025]. *Applied Sciences*, *13*(16), 9443. https://doi.org/10.3390/app13169443

Cigliano, A., Fallucchi, F., & Gerardi, M. (2024). The impact of digital analysis and large language models in digital humanity [[Online]. Accessed: 18 July 2025]. *ICYRIME 2024: 9th International Conference of Yearly Reports on Informatics, Mathematics, and Engineering*, *3869*, 1–6. https://ceur-ws.org/Vol-3869/p01.pdf

Didehkhani, Z. (2024). Analyzing persian twitter sentiments on the arbaeen walk: A comparative study of lda and bertopic with the arbaeen tweets dataset [[Online]. Accessed: 18 July 2025].

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts [[Online]. Accessed: 18 July 2025]. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Gan, L., Lu, H., & Cai, J. (2024). Experimental comparison of three topic modeling methods with lda, top2vec and bertopic [[Online]. Accessed: 18 July 2025]. In *Artificial intelligence and robotics* (pp. 376–391, Vol. 1998). https://doi.org/10.1007/978-981-99-9109-9_37

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure [[Online]. Accessed: 18 July 2025]. https://doi.org/10.48550/arXiv.2203.05794

Hosseiny Marani, A., & Baumer, E. P. S. (2023). A review of stability in topic modeling: Metrics for assessing and techniques for improving stability [[Online]. Accessed: 18 July 2025]. *ACM Computing Surveys*, *56*(5), 108:1–108:32. https://doi.org/10.1145/3623269

Khodeir, N., & Elghannam, F. (2025). Efficient topic identification for urgent mooc forum posts using bertopic and traditional topic modeling techniques [[Online]. Accessed: 18 July 2025]. *Education and Information Technologies*, *30*(5), 5501–5527. https://doi.org/10.1007/s10639-024-13003-4

Kumar, A., Karamchandani, A., & Singh, S. (2024). Topic modeling of neuropsychiatric diseases related to gut microbiota and gut brain axis using artificial intelligence based bertopic model on pubmed abstracts [[Online]. Accessed: 18 July 2025]. *Neuroscience Informatics*, *4*(4), 100175. https://doi.org/10.1016/j.neuri.2024.100175

Li, H. e. a. (2023). Research on a data mining algorithm based on bertopic for medication rules in traditional chinese medicine prescriptions [[Online]. Accessed: 18 July 2025]. *Medicine Advances*, *1*(4), 353–360. https://doi.org/10.1002/med4.39

Li, X. e. a. (2025). Evaluation of unsupervised static topic models' emergence detection ability [[Online]. Accessed: 18 July 2025]. *PeerJ Computer Science*, *11*, e2875. https://doi.org/10.7717/peerj-cs.2875

Lin, Q. e. a. (2024). Research on topic mining and evolution trends of functional agriculture based on the bertopic model [[Online]. Accessed: 18 July 2025]. *Agriculture*, *14*(10), 1691. https://doi.org/10.3390/agriculture14101691

Madrid-García, A. e. a. (2024). Mapping two decades of research in rheumatology-specific journals: A topic modeling analysis with bertopic [[Online]. Accessed: 18 July 2025]. *Therapeutic Advances in Musculoskeletal Disease*, *16*, 1759720X241308037. https://doi.org/10.1177/1759720X241308037

Medvecki, D. e. a. (2024). Multilingual transformer and bertopic for short text topic modeling: The case of serbian [[Online]. Accessed: 18 July 2025]. In *Disruptive information technologies for a smart society* (pp. 161–173, Vol. 872). https://doi.org/10.1007/978-3-031-50755-7_16

Ranaweera, N. e. a. (2025). Bertdetect: A neural topic modelling approach for android malware detection [[Online]. Accessed: 18 July 2025]. *Companion Proceedings of the ACM on Web Conference 2025*, 1802–1810. https://doi.org/10.1145/3701716.3717501

Wang, X. e. a. (2024). Digital deduction theatre: An experimental methodological framework for the digital intelligence revitalisation of cultural heritage [[Online]. Accessed: 18 July 2025]. In *Intelligent computing for cultural heritage* (pp. 203–220). https://library.oapen.org/bitstream/handle/20.500.12657/92133/9781040113264.pdf?sequence=1#page=232

Zhang, N., & Wang, J. (2024). Topic analysis of digital preservation based on bertopic [[Online]. Accessed: 18 July 2025]. *iPRES 2024*. https://doi.org/10.21428/5676bf2d.6f3ce886

Appendix A. Tables

Table 1: Overview of Relationship between Parameters and Error Size

min_cluster_size	min_topic_size	n_neighbors	error_size
10	30	5	6829
20	35	5	7275
10	50	5	7326
25	35	5	7353

Table 2: Evaluation Terms in Literature

Term	Frequency	Percentage
word embedding	33	61
Pointwise Mutual Information	11	20
cosine similarity	15	28
qualitative	23	43

Table 3: Topic Model Literature

Rank	Term	Frequency	Percentage
0	Latent Dirichlet Allocation	47	87%
7	UMAP	43	80%
2	HDBSCAN	40	74%
1	top2vec	32	59%
3	k-means	15	28%
5	min_cluster_size	5	9%
6	n_neighbors	5	9%
4	min_topic_size	4	7%

Table 4: Comparing BERTopic, Top2Vec and LDA

Metric	BERTopic	Top2Vec	LDA
PUV	0.968	0.926	0.947
Coherence (c_npmi)	-0.075	-0.270	-0.062
% Appearance	9.2%	15.2%	100.0%
Gini coefficient	0.139	0.114	0.140

Table 5: Large Corpus Descriptives

Index	Mean	Std. Dev.	Min	Max
Error_Size	56281.51	4604.76	41846.0	62180.0
Gini_Score	0.46	0.10	0.32	0.76

Table 6: Total BERTopic Runs

Run Name	**Total Values**	**Unique Values**	**Unique Percentage**
PUV (large)	342	55	16%
RUN (large)	301	41	14%
PUV (small)	405	113	28%
RUN (small)	92	92	100%
Total	**1140**	**301**	**26%**