IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3818~3826
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3818-3826

Genetic algorithm-optimized BERTopic with SHAP explainability for institutional research trend analysis

Muhammad Dedi Irawan[1], Yustria Handika Siregar[2], Hewa Majeed Zangana[3], Ali Ikhwan[4,5]
[1]Department of Data Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Deli Serdang, Indonesia
[2]Department of Computer Graphics Engineering Technology, Politeknik Cendana, Medan, Indonesia
[3]Department of Information Technology, Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq
[4]Department of Information Systems, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Deli Serdang, Indonesia
[5]Universiti Malaysia Perlis, Arau, Malaysia

Article Info

Article history:

Received Apr 11, 2025
Revised Jul 10, 2026
Accepted Jul 21, 2026

Keywords:

BERTopic
Genetic algorithm
Grey literature
Hyperparameter optimization
Semantic topic modeling
Shapley additive explanations
Short text analysis

ABSTRACT

Institutional research grant titles constitute short-text grey literature characterized by heterogeneous semantic structures, making topic identification and research trend analysis challenging. This study proposes an integrated bidirectional encoder representation from transformers-based topic modeling (BERTopic) framework combining genetic algorithm (GA)-based hyperparameter optimization and Shapley additive explanations (SHAP)-based interpretability to improve semantic topic quality and model transparency. GA was applied to optimize dimensionality-reduction and density-based clustering parameters, while SHAP was used to estimate the contribution of bigram features to the surrogate classifier's predictions of BERTopic-generated topic labels. Experimental results demonstrated that the proposed framework improved topic coherence from 0.367 to 0.543 while reducing the outlier ratio from 21.12% to 13.55%. In addition, the number of topics decreased from 46 to 10, resulting in a more compact and less fragmented topic structure. The resulting topic structure revealed dominant themes related to higher education, religious moderation, Islamic counseling, halal tourism, and sharia banking. Overall, the proposed framework contributes to the development of more coherent, interpretable, and semantically robust topic modeling for institutional short-text grey literature analysis.

Corresponding Author:

Muhammad Dedi Irawan
Department of Data Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara
Deli Serdang, North Sumatra 20353, Indonesia
Email: md.irawan@uinsu.ac.id

1. INTRODUCTION

The development of digital transformation has significantly increased the amount of institutional grey literature, including research grant titles, internal reports, and academic documents that are not published through formal scientific channels [1], [2]. In this study, grey literature is represented in the form of institutional research grant titles, which are characterized by short text, heterogeneity, and lack of standardization. The high semantic variation within these short texts makes topic identification and research trend analysis more challenging. In addition to limited linguistic context, imbalanced thematic distributions may also produce unstable topic structures that are difficult to interpret semantically.

The advancement of transformer-based natural language processing (NLP) has encouraged the emergence of semantic topic modeling approaches capable of understanding contextual relationships between

words and documents more effectively than traditional probabilistic approaches [3]–[5]. One rapidly growing method is bidirectional encoder representations from transformers-based topic modeling (BERTopic), which combines transformer embeddings, uniform manifold approximation and projection (UMAP) dimensionality reduction, hierarchical density-based spatial clustering of applications with noise (HDBSCAN) clustering, and class-based term frequency–inverse document frequency (c-TF-IDF) topic representation [6]. This approach is considered capable of generating more contextual topic representations and improving topic coherence across various research domains [7]–[9].

Several studies have applied BERTopic to aviation safety analysis, social media, the internet of things, and generative artificial intelligence trends [10]–[13]. These studies demonstrate that BERTopic performs well in capturing semantic structures and topic variations within complex datasets. However, BERTopic performance is highly dependent on the configuration of UMAP and HDBSCAN parameters. Suboptimal parameter selection may result in a high number of outliers, low topic coherence, and the formation of less representative clusters [7], [9]. This issue becomes even more complex when applied to grey literature datasets characterized by imbalanced data distributions and high heterogeneity.

To address these issues, this study applies a genetic algorithm (GA) as a hyperparameter optimization approach for BERTopic. GA possesses global exploration capabilities for identifying optimal parameter combinations within multidimensional search spaces, making it effective in improving clustering quality and machine learning model stability [14]–[16]. Although numerous studies have discussed hyperparameter optimization in machine learning models, the implementation of GA-based parameter optimization for BERTopic in grey literature analysis remains relatively limited. Beyond model quality, interpretability also presents a major challenge in embedding-based approaches because the topic formation process is often difficult to explain transparently.

The explainable artificial intelligence (XAI) approach based on Shapley additive explanations (SHAP) can be utilized to improve model interpretability by identifying feature contributions to prediction or topic structure formation [17], [18]. In the context of topic modeling, SHAP enables the identification of bigram feature contributions in forming dominant themes, thereby making the model interpretation process more transparent and understandable. The integration of GA-based parameter optimization and SHAP-based interpretability is expected to improve both the quality and transparency of BERTopic in analyzing grey literature datasets.

Based on these issues, this study proposes a BERTopic approach optimized using a GA and analyzed using SHAP for identifying research trends in grey literature. The objectives of this study are to: i) optimize UMAP and HDBSCAN parameters using GA, ii) evaluate improvements in topic quality based on topic coherence, topic diversity, and outlier ratio, and iii) analyze the contribution of bigram features to topic formation using SHAP. The contribution of this research lies in the integration of hyperparameter optimization and XAI-based interpretability within the BERTopic framework to produce a more stable, representative, and transparent topic modeling approach for heterogeneous and unstructured grey literature datasets.

2. METHOD

This study employs a GA-optimized BERTopic framework to improve topic quality and interpretability in institutional grey literature analysis. The framework integrates dataset collection, text preprocessing, sentence embedding generation, baseline BERTopic modeling, GA-based optimization of UMAP and HDBSCAN hyperparameters, optimized BERTopic modeling, and SHAP-based explainability. The overall research workflow is presented in Figure 1.

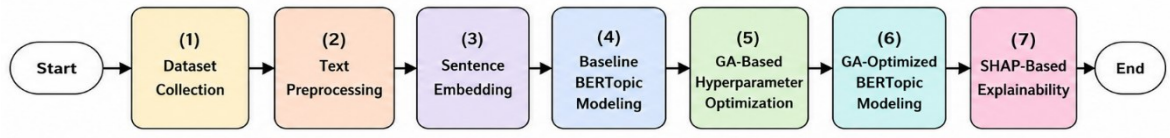

Figure 1. Proposed GA-optimized BERTopic framework with SHAP-based explainability

2.1. Dataset and preprocessing

The dataset comprised 502 research grant titles funded by Universitas Islam Negeri Sumatera Utara between 2018 and 2024. The data were obtained from the institutional research database managed by the Institute for Research and Community Service. Duplicate, incomplete, and administrative-only entries were

removed before analysis. The short-text dataset was categorized as academic grey literature and characterized by heterogeneous terminology and limited contextual information [1], [2]. Preprocessing included case folding, punctuation removal, tokenization, stopword removal, normalization, and bigram construction. Documents were then transformed into contextual embeddings using the all-MiniLM-L6-v2 SentenceTransformer model [4], [5], [19].

2.2. BERTopic framework

This study employed BERTopic to generate semantic topic structures using transformer embeddings, UMAP dimensionality reduction, HDBSCAN clustering, and c-TF-IDF topic representation [6], [20], [21]. UMAP was used to reduce embedding dimensions, while HDBSCAN grouped documents based on density distributions. Documents assigned to topic label −1 were treated as outliers. Topic representations were generated using c-TF-IDF as formulated in (1).

$$c\text{-}TF\text{-}IDF_{t,x} = \frac{f_{t,x}}{\sum_k f_{k,x}} \cdot log\left(\frac{N}{n_t}\right) \quad (1)$$

Where $f_{t,x}$ denotes the frequency of term t in topic x, N represents the total number of documents, and n_t indicates the number of topics containing term t.

2.3. Genetic algorithm optimization

A GA was applied to optimize UMAP and HDBSCAN parameters, including n_neighbors, min_cluster_size, and min_samples [14]–[16]. The search space consisted of n_neighbors ∈ {5, 8, 10, 12, 15}, min_cluster_size ∈ {5, 8, 10, 12, 15}, and min_samples ∈ {1, 2, 3, 5}. The UMAP parameters n_components and min_dist were fixed at 5 and 0.0, respectively. The optimization process used a population size of 5 and 3 generations with tournament selection, crossover probability of 0.8, and mutation probability of 0.1.

The fitness function combined topic coherence, topic diversity, outlier ratio, and topic compactness, as defined in (2).

$$Fitness = 0.35C + 0.25D + 0.25(1 - O) + 0.15T \quad (2)$$

Where C denotes the topic coherence, D represents the topic diversity, O indicates the outlier ratio, and T denotes the topic compactness. For each parameter combination, a BERTopic model was generated and evaluated iteratively to identify the optimal semantic topic structure with improved clustering stability and reduced topic fragmentation. The GA optimization process consisted of population initialization, fitness evaluation, selection, crossover, mutation, and population updates until the stopping criteria were achieved.

2.4. SHAP-based explainability

SHAP was applied to interpret the surrogate classifier by estimating the contribution of TF-IDF bigram features to the prediction of BERTopic-generated topic labels [17], [18]. Because BERTopic is unsupervised, topic labels generated by the optimized model were used in a surrogate logistic regression classifier with TF-IDF bigram features. SHAP LinearExplainer was then employed to estimate feature importance for semantic topic representation.

2.5. Model evaluation

Model performance was evaluated using topic coherence, topic diversity, outlier ratio, and number of topics [7], [20], [22]. Comparative analysis was conducted between the baseline BERTopic model and the GA-optimized BERTopic model. This comparison aimed to assess semantic coherence, clustering stability, and topic compactness.

3. RESULTS AND DISCUSSION

3.1. Evaluation of genetic algorithm-optimized BERTopic

The model performance was evaluated by comparing the baseline BERTopic model with default parameter settings and the GA-optimized BERTopic model using topic coherence, topic diversity, outlier ratio, and number of topics metrics, as presented in Table 1. The results in Table 1 indicate that parameter optimization using GA improved topic modeling quality. Topic coherence increased by approximately 48.0%, indicating enhanced semantic consistency among documents within each topic. In addition, the outlier ratio decreased by approximately 35.8%, demonstrating improved model capability in clustering

documents into more representative groups. The reduction in the number of topics from 46 to 10 also produced a more compact and stable topic structure. Although a slight decrease in topic diversity was observed, the value remained within a high range, suggesting that the trade-off for improved coherence was still acceptable. The GA optimization process identified an optimal parameter combination consisting of n_neighbors = 12, min_cluster_size = 15, and min_samples = 1. This parameter configuration produced the highest fitness score by improving semantic coherence while reducing topic fragmentation and clustering outliers in the short-text grey literature dataset.

Table 1. Performance comparison between baseline BERTopic and GA-Optimized BERTopic

Metric	Baseline	GA
Topic coherence	0.367	0.543
Topic diversity	~1.000	~0.988
Outlier ratio	21.12%	13.55%
Number of topics	46	10

3.2. Topic structure of genetic algorithm-optimized BERTopic

The semantic relationships among the topics generated by the GA-optimized BERTopic model were visualized using an intertopic distance map, as presented in Figure 2. The map was used to examine the relative semantic proximity and separation among the topics identified by the optimized model. As shown in Figure 2, the GA-optimized BERTopic model generated ten main topics, labeled T0–T9. Documents assigned to topic −1 were classified as outliers and were therefore excluded from the visualization. The map shows that several topics are positioned in close semantic proximity, indicating that they share related thematic content, whereas other topics occupy more distant positions and therefore exhibit relatively distinctive semantic characteristics.

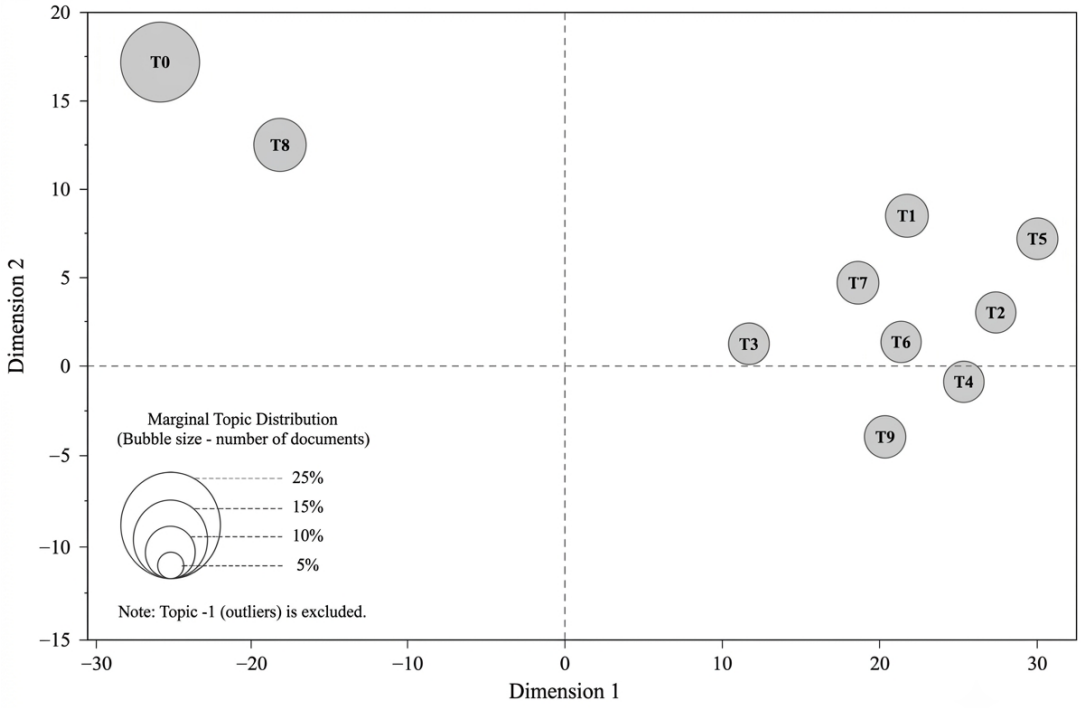

Figure 2. Intertopic distance map of the GA-optimized BERtopic model

The distribution presented in Figure 2 suggests that the GA-based optimization produced a compact topic structure with manageable semantic separation. The relatively close positioning of several topics should not necessarily be interpreted as poor clustering performance, because the institutional research corpus contains multidisciplinary themes that may naturally share related terminology and contextual characteristics.

At the same time, the more distant positions of several topics indicate the presence of themes with more distinctive semantic structures.

Overall, the optimized model reduced excessive topic fragmentation and produced a more interpretable structure consisting of ten main topics. The identified themes include higher education, religious moderation, social development, Islamic counseling, halal tourism, sharia banking, and several technical and legal domains. These findings indicate that the institutional grey literature corpus reflects multidisciplinary research trends with varying degrees of semantic interrelationship.

To strengthen the interpretation of the topic structure, the representative terms associated with the selected topics were further examined using c-TF-IDF scores. The distribution of the representative topic words is presented in Figure 3. The representative words with the highest c-TF-IDF scores across the ten topics generated by the GA-optimized BERTopic model. These terms characterize the semantic content of each topic and facilitate topic interpretation. The extracted keywords indicate that the model captured major themes related to higher education, religious moderation, Islamic counseling, halal tourism, sharia banking, machine learning, and other interdisciplinary research domains. Overall, the c-TF-IDF representations demonstrate that the optimized model produced semantically meaningful and distinguishable topic structures within the institutional research corpus.

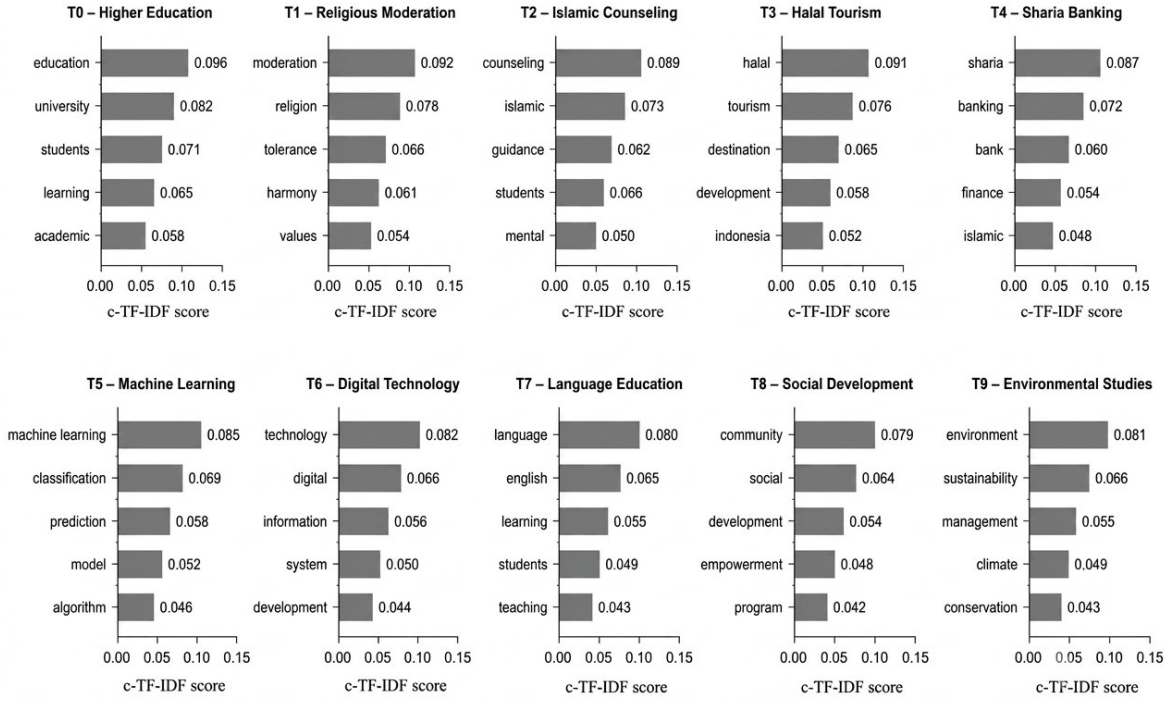

Figure 3. Representative topic word scores based on c-TF-IDF representation

3.3. Shapley additive explanations-based theme importance analysis

To improve model interpretability, this study applied SHAP to estimate the contribution of bigram features to the prediction of topic labels generated by the GA-optimized BERTopic model through a surrogate classifier. Figure 4 shows that higher education had the highest contribution, with a mean absolute SHAP value of approximately 0.044, followed by religious moderation and Islamic counseling. The dominance of these themes indicates that education- and religion-related issues played a major role in the semantic structure identified within the institutional grey literature corpus. Other influential bigram features included halal tourism, early childhood, sharia banking, machine learning, community development, public policy, artificial intelligence, and deep learning. The presence of these features reflects the interdisciplinary nature of the dataset, encompassing educational, religious, social, economic, policy, and technological domains.

The comparatively low contribution of convolutional neural network (CNN) algorithm suggests that this feature had only a limited influence on the surrogate model's prediction of BERTopic-generated topic labels. Its appearance may indicate residual semantic noise or a highly specific technical expression

occurring in a relatively small subset of documents. Overall, the SHAP analysis provides a more transparent interpretation of the features associated with the topic-label predictions and strengthens the explainability of the GA-optimized BERTopic framework.

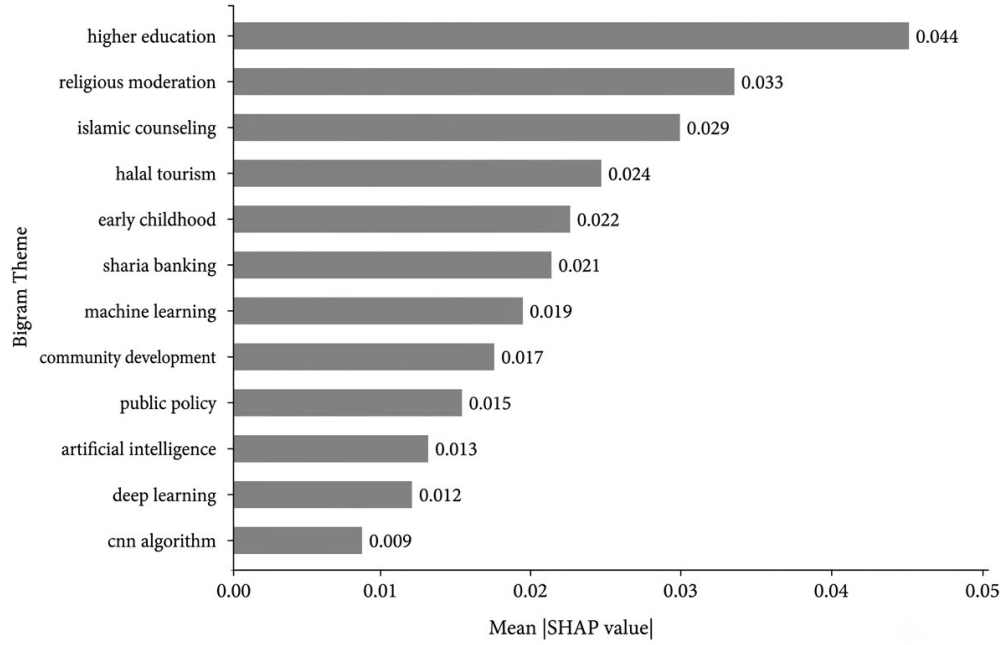

Figure 4. SHAP bigram theme importance for the GA-optimized BERtopic model

3.4. Discussion

The findings demonstrate that optimizing UMAP and HDBSCAN parameters using a GA improved topic modeling quality in grey literature datasets consisting of institutional research grant titles. Topic coherence increased from 0.367 to 0.543, while the outlier ratio decreased from 21.12% to 13.55%, indicating that the optimization process successfully generated a more stable and representative clustering structure compared to the baseline BERTopic model. These findings suggest that parameter configuration significantly influences semantic representation quality in BERTopic, particularly for short-text datasets characterized by high sparsity and imbalanced topic distributions [20], [21].

The reduction in the number of topics from 46 to 10 indicates that GA-based optimization successfully reduced excessive topic fragmentation while improving semantic consistency among documents. In short-text topic modeling, topic fragmentation frequently occurs due to limited linguistic context within each document, making semantic relationships between texts more difficult to preserve [21]. Compared conceptually with conventional probabilistic topic modeling approaches such as latent Dirichlet allocation (LDA) and embedding-based methods such as Top2Vec, the proposed approach produced more compact and interpretable topic representations through the integration of transformer embeddings and evolutionary computation-based optimization [23]. These findings suggest that transformer embeddings combined with evolutionary optimization may improve semantic consistency and clustering stability in short-text grey literature datasets compared to approaches without parameter optimization.

The intertopic distance map visualization in Figure 2 shows that most topics exhibit relatively good semantic separation with minimal overlap between clusters. These results indicate that GA-based optimization of UMAP and HDBSCAN parameters effectively preserved the semantic structure of document embeddings. These findings are consistent with previous studies stating that clustering quality in BERTopic is highly influenced by dimensionality reduction configuration and document embedding distribution [20], [22].

In addition to improving clustering quality, this study also demonstrates that the SHAP approach enhanced BERTopic model transparency by identifying the contribution of bigram features to topic formation. The dominance of themes such as higher education, religious moderation, and Islamic counseling indicates that the model successfully captured the primary semantic structures within the institutional research grant dataset. These findings suggest that educational and religious moderation issues remain dominant themes in the development of institutional research within grey literature. Furthermore, the emergence of multidisciplinary themes such as halal tourism, sharia banking, and machine learning indicates

that institutional research topics are no longer confined to a single domain, but have evolved toward the integration of educational, social, economic, and technological issues. These findings are consistent with previous studies reporting that BERTopic and transformer-based topic modeling approaches are effective in extracting semantically coherent topics from unstructured and short-text datasets across diverse application domains, including social media, public discourse, healthcare, and religious studies [4], [19], [23]–[26].

Nevertheless, several low-contribution bigrams such as CNN algorithm and best palm still appeared in the SHAP results. The presence of these features indicates residual semantic noise in certain documents caused by terminology variation and the limited context of short texts. However, the influence of these features was relatively minor compared to the dominant themes and therefore did not significantly affect the overall topic structure. This condition demonstrates that GA-based optimization improved model stability without eliminating thematic diversity within the dataset. Overall, the integration of GA-based hyperparameter optimization and SHAP-based interpretability provides a methodological contribution to the development of semantic topic modeling for short-text grey literature datasets. The proposed approach not only improved topic representation quality but also produced more transparent and interpretable model explanations for institutional research trend analysis.

Despite the promising results, this study has several limitations. First, the proposed framework was evaluated only on a single institutional short-text grey literature dataset consisting of research grant titles. Second, direct experimental benchmarking against other topic modeling approaches such as LDA and Top2Vec was not conducted. Third, stakeholder-based qualitative validation was not included because the study primarily focused on quantitative evaluation metrics and semantic interpretability. Future studies may address these limitations through cross-dataset evaluation, multilingual corpora, full-text document analysis, and integration with large language models (LLMs) to further improve semantic topic representation and model generalizability.

4. CONCLUSION

This study proposed a BERTopic approach optimized using a GA and interpreted using SHAP for trend analysis on institutional short-text grey literature. The experimental results demonstrate that optimizing UMAP and HDBSCAN parameters using GA significantly improved topic modeling quality. Topic coherence increased from 0.367 to 0.543, while the outlier ratio decreased from 21.12% to 13.55%. In addition, the number of topics was reduced from 46 to 10, resulting in a more compact, stable, and semantically interpretable topic structure. The visualization of the intertopic distance map and c-TF-IDF-based word distributions showed that the model was capable of capturing semantic relationships among topics with good separability. The dominant themes identified included higher education, religious moderation, Islamic counseling, halal tourism, and sharia banking, reflecting the development of multidisciplinary themes in institutional research. SHAP-based analysis also demonstrated that bigram features contributed to the surrogate model's prediction of BERTopic-generated topic labels, thereby improving model transparency and interpretability. Overall, the integration of GA-based hyperparameter optimization and SHAP-based interpretability provides a methodological contribution to the development of semantic topic modeling for short-text grey literature datasets. The proposed approach not only improves topic representation quality but also produces more transparent interpretations in institutional research trend analysis. Future studies may extend this approach to multilingual datasets, full-text documents, or integration with LLMs to further enhance semantic representation quality in a more comprehensive manner.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Institute for Research and Community Service (LPPM), Universitas Islam Negeri Sumatera Utara, for its institutional support.

FUNDING INFORMATION

This research was funded by Universitas Islam Negeri Sumatera Utara through the 2024 DIPA High-er Education Development Research Grant (Agreement No. B.24/Un.11.R/L2.3/KS/0.2/03/2024).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Dedi Irawan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Yustria Handika Siregar				✓		✓	✓	✓		✓	✓		✓	
Hewa Majeed Zangana		✓		✓	✓					✓		✓		
Ali Ikhwan	✓		✓						✓					✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

INFORMED CONSENT

Informed consent is not applicable because this study does not involve human participants or personal data.

ETHICAL APPROVAL

Ethical approval was not required because this study used institutional research grant title data and did not involve human subjects or sensitive personal information.

DATA AVAILABILITY

The dataset used in this study is available from the corresponding author, [MDI], upon reasonable request.

REFERENCES

[1] A. Paez, "Gray literature: an important resource in systematic reviews," *Journal of Evidence-Based Medicine*, vol. 10, no. 3, pp. 233–240, Aug. 2017, doi: 10.1111/jebm.12266.

[2] J. A. Wani and S. A. Ganaie, "The scientific outcome in the domain of grey literature: bibliometric mapping and visualisation using the R-bibliometrix package and the VOSviewer," *Library Hi Tech*, vol. 42, no. 1, pp. 309–330, Feb. 2024, doi: 10.1108/LHT-01-2022-0012.

[3] S. Mifrah and E. H. Benlahmar, "Topic modeling with transformers for sentence-level using coronavirus corpus," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 17, pp. 50–59, Sep. 2022, doi: 10.3991/ijim.v16i17.33281.

[4] A. Zafar, M. Wasim, S. Zulfiqar, T. Waheed, and A. Siddique, "Transformer-based topic modeling for urdu translations of the holy quran," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 10, pp. 1–21, Oct. 2024, doi: 10.1145/3694967.

[5] M. A. Khadija and W. Nurharjadmo, "Enhancing Indonesian customer complaint analysis: LDA topic modelling with BERT embeddings," *SINERGI*, vol. 28, no. 1, pp. 153–162, Dec. 2023, doi: 10.22441/sinergi.2024.1.015.

[6] M. Grootendorst, "BERTopic: neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, *arXiv:2203.05794*.

[7] C. Yang and Y. Kim, "Enhancing topic coherence and diversity in document embeddings using LLMs: a focus on BERTopic," *Expert Systems with Applications*, vol. 281, Jul. 2025, doi: 10.1016/j.eswa.2025.127517.

[8] D. Medvecki, B. Bašaragin, A. Ljajić, and N. Milošević, "Multilingual transformer and BERTopic for short text topic modeling: the case of Serbian," in *Conference on Information Technology and its Applications*, 2024, pp. 161–173, doi: 10.1007/978-3-031-50755-7_16.

[9] D. P. Gottumukkala, P. R. P.V.G.D, and S. K. Rao, "Topic modeling-based prediction of software defects and root cause using BERTopic, and multioutput classifier," *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-11458-0.

[10] A. Nanyonga, K. Joiner, U. Turhan, and G. Wild, "Semantic topic modeling of aviation safety reports: a comparative analysis using BERTopic and PLSA," *Aerospace*, vol. 12, no. 6, Jun. 2025, doi: 10.3390/aerospace12060551.

[11] H. Son and Y. E. Park, "Agenda-setting effects for COVID-19 vaccination: insights from 10 million textual data from social media and news articles using BERTopic," *International Journal of Information Management*, vol. 83, Aug. 2025, doi: 10.1016/j.ijinfomgt.2025.102907.

[12] Y.-S. Su, J. Wang, S.-H. Tu, K.-T. Liao, and C.-L. Lin, "Detecting latent topics and trends in IoT and e-commerce using BERTopic modeling," *Internet of Things*, vol. 32, Jul. 2025, doi: 10.1016/j.iot.2025.101604.

[13] M. Kilinc, F. Gurcan, and A. Soylu, "LLM-based generative AI in medicine: analysis of current research trends with BERTopic," *IEEE Access*, vol. 13, pp. 157026–157043, 2025, doi: 10.1109/ACCESS.2025.3606335.

[14] H. R.-Berumen, A. Zafra, and S. Ventura, "A survey of genetic algorithms for clustering: taxonomy and empirical analysis," *Swarm and Evolutionary Computation*, vol. 91, Dec. 2024, doi: 10.1016/j.swevo.2024.101720.

[15] B. Bischl *et al.*, "Hyperparameter optimization: foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, Mar. 2023, doi: 10.1002/widm.1484.

[16] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, Nov. 2020, doi: 10.1016/j.neucom.2020.07.061.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, pp. 4768–4777, 2017.
[18] R. Muthusami, K. Saritha, K. S. Rao, P. Sugapriya, and G. Saveetha, "Interpretable stance detection in social media via topic-guided transformers," *Discover Artificial Intelligence*, vol. 5, no. 1, Nov. 2025, doi: 10.1007/s44163-025-00635-9.
[19] L. Ma *et al.*, "AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women," *Experimental Biology and Medicine*, vol. 250, Feb. 2025, doi: 10.3389/ebm.2025.10389.
[20] H. Okazaki and H. Takahashi, "Analysis of extracting corporate trends from news data using BERTopic," *New Generation Computing*, vol. 43, no. 1, Feb. 2025, doi: 10.1007/s00354-025-00290-9.
[21] K.-K. Lai, C.-W. Hsiao, and Y.-J. Hsu, "Strategic management knowledge map via BERTopic (1980–2025): evolution, integration, and application," *Applied System Innovation*, vol. 8, no. 5, Sep. 2025, doi: 10.3390/asi8050142.
[22] A. Nanyonga, K. Joiner, U. Turhan, and G. Wild, "Does the choice of topic modeling technique impact the interpretation of aviation incident reports? a methodological assessment," *Technologies*, vol. 13, no. 5, May 2025, doi: 10.3390/technologies13050209.
[23] A. Kaur and J. Wallace, "From transformers come themes: evaluating BERTopic for qualitative analysis of social media data," in *2025 International Conference on Human-Engaged Computing*, Nov. 2025, pp. 1–14, doi: 10.1145/3786995.3786997.
[24] R. Espinoza *et al.*, "Leveraging social media for public health: NLP implementations for blood donation data analysis in Japan," *Social Network Analysis and Mining*, vol. 15, no. 1, Apr. 2025, doi: 10.1007/s13278-025-01437-8.
[25] A. Alamsyah and R. F. Astuti, "Analyzing public discourse on DeFi and CBDC using advanced NLP techniques: insights for financial policy and innovation," *Digital Policy, Regulation and Governance*, vol. 27, no. 6, pp. 703–723, Oct. 2025, doi: 10.1108/DPRG-09-2024-0240.
[26] R. Egger and J. Yu, "A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts," *Frontiers in Sociology*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

BIOGRAPHIES OF AUTHORS

Muhammad Dedi Irawan is a lecturer at the Department of Data Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Indonesia. He received bachelor's and master's degrees in Informatics from Universitas Asahan and Universitas Putra Indonesia YPTK Padang, respectively. His research interests include artificial intelligence, decision support systems, information technology, and computer science. He has published various scientific papers in computing and information systems and actively participates in academic research and community service projects. He has also been involved in nationally funded research activities. He can be contacted at email: md.irawan@uinsu.ac.id.

Yustria Handika Siregar is a lecturer in the Computer Graphic Engineering Program at Politeknik Cendana, Indonesia. She received bachelor's and master's degrees in Informatics from Universitas Asahan and Universitas Putra Indonesia YPTK Padang, respectively. Her research interests include artificial intelligence, decision support systems, image processing, and computing technology. She has contributed to several scientific publications in information technology and has been actively involved in research and community service activities. She can be contacted at email: yustriahandikasiregar@gmail.com.

Hewa Majeed Zangana is an assistant professor at Department of Information Technology, Duhok Technical College, Duhok Polytechnic University, Iraq, and currently pursuing a Ph.D. in Department of Information Technology Management at Duhok Polytechnic University. His research interests include network systems, information security, mobile communication, intelligent systems, and deep learning applications. He has held several academic and administrative positions at Duhok Polytechnic University and Nawroz University. He has contributed to scientific publications in cybersecurity, image processing, and artificial intelligence research. He can be contacted at email: hewa.zangana@dpu.edu.krd.

Ali Ikhwan He earned his bachelor's degree in Information Systems from STMIK Triguna Dharma, Medan, Indonesia, in 2013. He received his master's degree in Computer Science from Universitas Putra Indonesia YPTK, Padang, Indonesia, in 2015. He is currently pursuing a doctoral degree at Universiti Malaysia Perlis, Malaysia, with a research focus on data mining and cryptography. He is a lecturer at the Department of Information Systems, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Indonesia. His research interests include data analysis, data mining, and cryptography. He can be contacted at email: ali_ikhwan@uinsu.ac.id.