

# Integrating IndoBERT and balanced iterative reducing and clustering using hierarchies of BERTopic in Indonesian short text

Muhammad Muhajir<sup>1,2</sup>, Gunardi<sup>1</sup>, Danardono<sup>1</sup>, Dedi Rosadi<sup>1,3</sup>

<sup>1</sup>Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada, Sleman, Indonesia

<sup>2</sup>Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Sleman, Indonesia

<sup>3</sup>Statistics RnD, Jakarta, Indonesia

## Article Info

### Article history:

Received Jun 19, 2024

Revised Jul 11, 2025

Accepted Aug 6, 2025

### Keywords:

BERTopic

BIRCH

BM25

IndoBERT

Indonesian short text

## ABSTRACT

Short text topic modeling remains challenging due to data sparsity, limited word co-occurrences, and unstable clustering results, particularly for Indonesian texts. This study proposes an improved BERTopic framework that integrates IndoBERT embeddings, best match 25 (BM25)-based topic representation, and balanced iterative reducing and clustering using hierarchies (BIRCH) clustering to address these issues. IndoBERT generates contextual embeddings adapted to Indonesian linguistic features, and BM25 weighting improves keyword relevance by considering document length and term saturation. BIRCH clustering minimizes outliers by assigning most documents to valid clusters, which enhances data utilization and topic stability. Experiments on Indonesian datasets from X (formerly Twitter), Google Reviews, and YouTube demonstrate that the proposed approach consistently achieves higher topic coherence. The proposed method yields stable topic diversity values between 0.91 and 0.94, maintains embedding density from 0.60 to 0.66, and achieves intra-topic similarity between 0.39 and 0.41 across increasing dataset sizes. The proposed framework successfully reduces outlier proportions to 1-5%, which significantly outperforms standard BERTopic and K-Means. Furthermore, the model maintains stable topic counts as the data volume grows, confirming robustness and scalability for sparse short text modeling. Overall, integrating IndoBERT, BM25, and BIRCH provides a more coherent, stable, and effective solution for Indonesian short text topic modeling.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Muhammad Muhajir

Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada  
Bulaksumur, Caturtunggal, Depok, Sleman Regency, Special Region of Yogyakarta 55281, Indonesia

Email: muhammadmuhajir@mail.ugm.ac.id

## 1. INTRODUCTION

Topic modeling is a fundamental technique in natural language processing (NLP) used to discover hidden thematic structures within a collection of documents. It helps extract meaningful insights by identifying patterns of word co-occurrence, allowing large volumes of text data to be summarized into interpretable topics [1], [2]. Over the years, numerous topic modeling approaches have been developed, each attempting to balance accuracy, scalability, and interpretability.

The early development of topic modeling began with latent semantic analysis (LSA), originally known as latent semantic indexing (LSI), which used singular value decomposition to capture the latent semantic structure of textual data [3]. While LSA could uncover certain conceptual relationships between

terms, it struggled with synonymy and polysemy since it relied purely on linear algebraic operations without incorporating the actual meaning of words. To address these limitations, probabilistic latent semantic analysis (PLSA) was introduced, framing topic modeling within a probabilistic context where documents are treated as mixtures of topics, and words are generated according to these topic distributions [4]. Despite improving the flexibility of topic modeling, PLSA still used the bag-of-words assumption, neglecting word order and contextual meaning.

A major milestone was achieved with the development of latent dirichlet allocation (LDA), which formalized topic generation as a hierarchical Bayesian process. LDA remains one of the most widely used topic modeling algorithms due to its theoretical elegance and interpretability [5]. However, LDA and its probabilistic predecessors encounter significant challenges when applied to short texts, such as tweets, social media comments, or short online reviews. The limited length of these documents reduces the likelihood of word co-occurrences, making it difficult for these models to infer meaningful topics. In short texts, data sparsity results in unstable topic distributions and poor coherence [6], [7].

To address the data sparsity issue in short texts, biterm topic modeling (BTM) was proposed. Unlike LDA, which models topics within each document, BTM analyzes word pairs (biterms) across the entire corpus, allowing it to better capture global word co-occurrence patterns [8]. This corpus-level modeling of biterms helps reduce the sparsity problem and improves topic coherence for short text data. Nonetheless, while BTM alleviates some limitations, it still primarily depends on surface-level co-occurrence patterns and lacks the capacity to model complex semantic relationships between words [9].

The recent advances in deep learning and pre-trained language models have opened new avenues for topic modeling. One such approach is BERTopic, which leverages embeddings produced by the bidirectional encoder representations from transformers (BERT) language model [10]. BERTopic creates dense semantic representations of documents at the sentence level, capturing contextual meaning beyond simple co-occurrences. These embeddings are then reduced in dimensionality using uniform manifold approximation and projection (UMAP), preserving both global and local data structures [11], and finally clustered to identify topics. The class-based term frequency-inverse document frequency (c-TF-IDF) method is then applied to extract representative keywords for each cluster [12].

While BERTopic has demonstrated strong performance across many domains, it still presents notable limitations when applied to short texts. The standard clustering algorithm used in BERTopic, hierarchical density-based spatial clustering of applications with noise (HDBSCAN), often identifies a large number of outliers in short text data, discarding valuable information. Empirical studies have shown that HDBSCAN may classify up to 74% of short text documents as outliers, significantly limiting topic coverage and stability. Moreover, c-TF-IDF struggles to fully capture semantic relationships, particularly in sparse, noisy datasets, and can generate less coherent topics when word frequency distributions are highly skewed [13].

To address these shortcomings, this research proposes several modifications to improve the performance of BERTopic on Indonesian short text datasets. First, we modify the clustering process by replacing HDBSCAN with balanced iterative reducing and clustering using hierarchies (BIRCH), a hierarchical clustering algorithm known for its memory efficiency and robustness to outliers [13], [14]. Unlike K-Means, which requires a predefined number of clusters and is sensitive to initialization, BIRCH dynamically forms clusters while minimizing memory usage and is better suited for large datasets with varying cluster shapes [15].

Second, we incorporate the best match 25 (BM25) ranking function to replace c-TF-IDF in the topic representation stage. BM25, widely used in information retrieval, better accounts for term frequency saturation and document length normalization, allowing more meaningful weighting of keywords [16], [17]. Additionally, we apply a frequency-based word reduction strategy to remove overly common terms that might dominate the topic representation, thus improving coherence by emphasizing more informative words.

Finally, for the embedding process, we adopt IndoBERT, a pre-trained BERT model specifically trained on Indonesian language corpora, including social media data such as Indonesian X (formerly Twitter) [18]. This adaptation allows the model to better understand the unique linguistic characteristics, colloquialisms, and domain-specific vocabulary prevalent in Indonesian short texts. IndoBERT produces high-quality contextualized embeddings that enhance the semantic understanding required for accurate topic modeling.

The contributions of this study are as follows. First, we introduce BIRCH clustering into the BERTopic framework to improve memory efficiency and address the outlier problem inherent in HDBSCAN. Second, we utilize BM25-based weighting and reduce the influence of highly frequent words to enhance semantic capture in topic representation. Third, we employ IndoBERT embeddings to better process and understand short Indonesian texts, which often present unique challenges compared to general multilingual datasets. Collectively, these modifications provide a more stable, accurate, and contextually relevant topic modeling framework for analyzing Indonesian short texts from diverse sources such as X, Google Reviews, and YouTube comments.

The structure of this investigation is outlined as follows. Section 1 discusses the background and significance of the study. Section 2 provides a brief overview of the improvements made to BERTopic through the integration of IndoBERT and BIRCH clustering. Section 3 presents a study based on Indonesian short text. Lastly, section 4 provides the conclusions.

## 2. METHOD

### 2.1. BERTopic

In this study, the BERTopic framework is modified to optimize topic modeling on Indonesian short-text data by incorporating four key components: document embedding, dimensionality reduction, and topic representation. First, document embedding is performed using IndoBERT, a BERT-based pre-trained language model specifically adapted for the Indonesian language. IndoBERT is capable of generating contextualized vector representations by considering bidirectional context for each word within a sentence [18]. The embedding process involves several stages, including preprocessing, tokenization, input representation, transformer layers, and bidirectional embeddings, which capture the full contextual meaning of each word. These embeddings are particularly effective for grouping documents with similar semantic meanings, though primarily utilized for semantic representation rather than directly generating topics.

Second, since the embedding vectors generated by IndoBERT are high-dimensional, a dimensionality reduction step is required to project the data into a lower-dimensional space while preserving its essential structure. Prior research has shown that, in high-dimensional data spaces, distances between data points tend to become increasingly uniform, making it difficult to distinguish semantic proximity [19]. To address this, dimensionality reduction techniques such as UMAP are employed. Compared to other techniques like PCA or t-distributed stochastic neighbor embedding (t-SNE), UMAP constructs a more representative low-dimensional graph while preserving the higher-order structure of the original data, where each node represents one document [11]. Furthermore, UMAP handles clusters with varying densities by calculating local radii for each data point, allowing more accurate topic clustering within BERTopic [20].

Third, after clustering is performed, topic representation is obtained using the BM25 algorithm. BM25 serves as an enhancement over the conventional term frequency-inverse document frequency (TF-IDF) algorithm by calculating relevance scores for each word relative to a topic, considering both term frequency and document length. This weighting approach accounts for not only word frequency but also the positional and proportional contribution of terms within the document, ensuring that highly informative keywords receive greater weight [21]. Overall, the application of BM25 in BERTopic enables more precise capture of semantic meaning and contextual relationships among words, resulting in more accurate, stable, and informative topic representations.

### 2.2. Improved BERTopic

The method proposed in this study is a modification of BERTopic using IndoBERT for sentence embedding in documents. In brief, sentence-level embeddings are generated for every document before computing their average to create a vector representation using UMAP, which reduces the dimensions further. Then, clustering is performed using the BIRCH algorithm with computed c-TF-IDF scores, aligning the classes and creating HDBSCAN vector representations. The BERTopic process is mathematically formulated as follows:

- i) Embed document:
  - Suppose  $A$  and  $B$  are given as  $A = [S_1^A, S_2^A, \dots, S_n^A]$  and  $B = [S_1^B, S_2^B, \dots, S_n^B]$ , where  $S_i^A$  and  $S_i^B$  are the sentences in the text, and  $n$  is the number of sentences.
  - Text  $A$  and  $B$  are embedded into  $E_A = [e_1^A, e_2^A, \dots, e_n^A]$  and  $E_B = [e_1^B, e_2^B, \dots, e_n^B]$  using IndoBERT [18], [22].
- ii) Dimensional reduction:
  - Take an embedding vector matrix  $X \in \mathbb{R}^{N \times M}$ , where  $N$  is the number of short texts, and  $M$  is the original dimension.
  - UMAP is used to reduce the dimension to  $Y \in \mathbb{R}^{N \times d}$ , where  $d$  is the target dimension [11].
- iii) BIRCH clustering:
  - Take a distance matrix  $D' \in \mathbb{R}^{N \times N}$  based on  $Y$ .
  - BIRCH clusters the data into  $C = \{C_1, C_2, \dots, C_k\}$  based on several parameters, such as threshold (T) and branching factor (B) [23], [24].
- iv) Topic representation:

- For each  $C_k$ , calculate the BM25 of the keyword:  $w_{t,c_k} = BM25(t, C_k) = \sum_{d \in C_k} IDF(t) \cdot \frac{f(t,d) \cdot (k_1+1)}{f(t,d) + k_1(1-b+b \cdot \frac{|d|}{avgdl})}$ , where  $f(t, d)$  is the frequency of term  $t$  in document  $d$ ,  $|d|$  is the document length,  $avgdl$  is the average document length in  $C$ ,  $k_1$  and  $b$  are adjusted parameters, and  $IDF(t) = \log(\frac{N-n(t)+0.5}{n(t)+0.5})$  [25].

### 2.3. Measuring topic similarity

To evaluate the quality of the generated topics, this study employs topic diversity, embedding density, and intra-topic similarity metrics. The selection of these metrics is motivated by the complexity of short-text characteristics, which are often difficult to assess using only conventional coherence measures [26], [27]. Topic diversity measures the extent to which the model produces distinct and non-redundant topics, thus reflecting its ability to capture various information aspects from the corpus [28]. Embedding density evaluates the compactness of embeddings within each cluster, where higher density indicates greater semantic consistency among documents within the same topic [29]. Meanwhile, intra-topic similarity calculates the average semantic similarity among sentences or documents within a single cluster, ensuring that each topic exhibits internal homogeneity and minimizes semantic ambiguity [30], [31]. Collectively, these metrics provide a more comprehensive assessment of the model's stability, clarity, and topical diversity.

### 2.4. Proposed method

The overall framework of the proposed modified BERTopic method is illustrated in Figure 1. It summarizes the key stages involved, starting from document embedding using IndoBERT and dimensionality reduction via UMAP. Clustering is then performed with BIRCH, followed by topic representation using BM25.

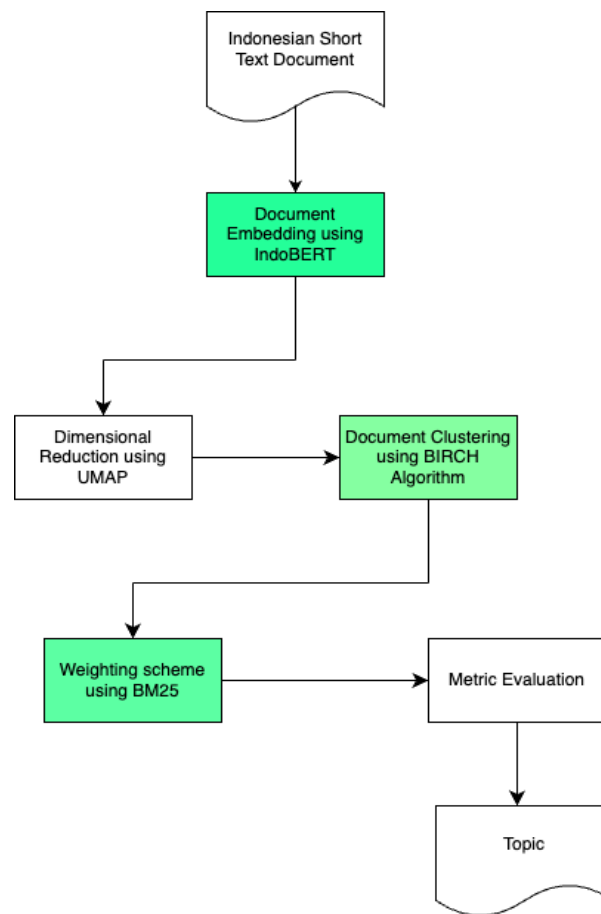


Figure 1. Framework of the proposed modified BERTopic method integrating IndoBERT, UMAP, BIRCH clustering, and BM25-based topic representation

3. RESULTS AND DISCUSSION

3.1. Experiment

This research uses Python’s text crawling feature to collect data. The first source includes reviews and comments from Indonesian X related to Hacker Bjorka, using the hashtags “Bjorkanisme” and “Hacker Bjorka,” resulting in a total of 3,061 collected comments. The second data source is Google Reviews on the Play Store, using the reviews of X apps, providing more than 11,800 user ratings for analysis. Finally, 46,896 comments were collected from users responding to the movie called dirty vote (depicting election fraud in Indonesia) on YouTube.

Preprocessing was performed using various techniques, including case folding, case elimination, hashtag removal, URL deletion, punctuation removal, word normalization, and stopword elimination using the sastrawi package. Specific words were also excluded to minimize cumulative discrepancies among verb tenses and ensure a clear outcome. The use of complex terminology and word stemming was avoided because their inclusion may increase the discrepancies among different verb tenses.

3.2. Improved BERTopic

Table 1 compares the topic modeling performance of standard BERTopic (HDBSCAN), BERTopic with K-Means, and the improved BERTopic with BIRCH. The comparison is conducted across X, Google Reviews, and YouTube datasets. The evaluation uses topic diversity, embedding density, and intra-topic similarity to assess the model’s ability to generate coherent, distinct, and meaningful topics from Indonesian short texts.

Table 1. Performance comparison of standard BERTopic, BERTopic integrated with K-Means clustering and BERTopic integrated with BIRCH clustering

Method	Dataset	Subset of documents (%)	Diversity	Density	Intra-similarity
BERTopic	X	20	0.9895	0.7448	0.4867
		40	0.9942	0.733	0.4599
		60	0.9953	0.7334	0.4517
		80	0.9964	0.7268	0.4442
		100	0.9967	0.7393	0.4808
BERTopic with K Means clustering	Google Reviews	20	0.8689	0.6083	0.369
		40	0.8239	0.5677	0.3523
		60	0.8229	0.5574	0.3709
		80	0.8111	0.5498	0.4052
		100	0.5839	0.6773	0.4897
BERTopic with BIRCH clustering	YouTube	20	0.9226	0.6426	0.3998
		40	0.9234	0.6092	0.3737
		60	0.9213	0.6109	0.4023
		80	0.911	0.6052	0.406
		100	0.9393	0.6589	0.5039

Table 1 presents the detailed evaluation of BERTopic performance across different clustering configurations (HDBSCAN, K-Means, and BIRCH) on three Indonesian short text datasets: X, Google Reviews, and YouTube. The experiments were conducted progressively on four data partitions (20%, 40%, 60%, and 80% subsets), allowing assessment of model stability and scalability as dataset sizes increase. The evaluation focuses on four key quality metrics: topic diversity, silhouette score, embedding density, and intra-topic similarity.

On the YouTube dataset, which contains highly noisy and diverse user-generated content, the improved BERTopic with BIRCH consistently demonstrates superior performance across all data partitions. As the dataset grows from 20 to 80%, topic diversity remains highly stable in the range of 0.922 to 0.911, indicating well-separated and distinct topic groups regardless of data size. Embedding density values for BIRCH remain steady between 0.642 and 0.605, reflecting compact cluster formation even as more documents are incorporated. Intra-topic similarity under BIRCH also maintains consistent values, ranging from 0.399 to 0.406, suggesting coherent topic groupings across subsets. In contrast, both standard BERTopic and K-Means exhibit greater fluctuation across subsets, with notably lower density and intra-similarity values, indicating more fragmented clustering as data volume increases.

For the Google Reviews dataset, which represents more structured and formal text, BERTopic with BIRCH again consistently outperforms other methods. The topic diversity achieved by BIRCH remains stable across subsets (0.922 to 0.911), while its embedding density shows gradual improvement from 0.642 to 0.605 as data volume increases, indicating efficient cluster compactness. Intra-topic similarity for

BIRCH remains consistently higher (around 0.40), while K-Means and HDBSCAN demonstrate more variability, with lower intra-topic similarity and less stable cluster structures as dataset size increases.

On the X dataset, which is characterized by highly dynamic and informal language, the superior capability of BIRCH is even more apparent. While standard BERTopic (HDBSCAN) suffers from extreme over-fragmentation at smaller data sizes (diversity reaching 0.989 at 20% subset), BIRCH maintains stable diversity levels between 0.922 and 0.911 across all subsets. Embedding density for BIRCH remains stable between 0.642 and 0.605 even as the data size grows, while K-Means and HDBSCAN exhibit declining density values, reflecting increasing cluster dispersion. Similarly, BIRCH maintains intra-topic similarity values around 0.40 across all partitions, which is consistently higher than both HDBSCAN and K-Means.

The progressive evaluations across subsets further confirm that the improved BERTopic with BIRCH clustering delivers not only superior topic coherence, but also higher model stability and scalability as more data becomes available. Unlike standard BERTopic or K-Means, which experience declining topic cohesion and cluster compactness as data grows, BIRCH consistently preserves both topic distinctiveness and intra-cluster coherence. This demonstrates that the proposed approach is highly suitable for real-world Indonesian short text applications, where data volume and variability often increase over time.

### 3.3. Outlier handling analysis

In addition to evaluating topic coherence, analyzing outlier proportions is essential to assess clustering effectiveness in short text data. Excessive outliers reduce data utilization and weaken topic representativeness. Table 2 summarizes the outlier percentages produced by BERTopic with HDBSCAN, K-Means, and BIRCH across the X, Google Reviews, and YouTube datasets.

Table 2. Comparison of outlier proportions across clustering methods on Indonesian short text datasets

Dataset	Clustering method	Documents classified as outliers (%)
X	BERTopic	23.15
X	BERTopic with K Means clustering	11.90
X	BERTopic with BIRCH clustering	3.77
Google Reviews	BERTopic	24.61
Google Reviews	BERTopic with K Means clustering	4.86
Google Reviews	BERTopic with BIRCH clustering	2.64
YouTube	BERTopic	23.97
YouTube	BERTopic with K Means clustering	6.66
YouTube	BERTopic with BIRCH clustering	4.84

The results in Table 2 clearly demonstrate the varying capabilities of the clustering algorithms in managing outliers for short text topic modeling. On the X dataset, the standard BERTopic using HDBSCAN classified approximately 23.15% of the documents as outliers. This high proportion aligns with previous findings that highlight HDBSCAN's tendency to struggle when dealing with sparse and fragmented short text data. While K-Means clustering significantly reduced the outlier proportion to 11.9%, a notable portion of documents still remained unassigned to any meaningful cluster. In contrast, BIRCH clustering demonstrated the best performance by minimizing outliers to only 3.77%, indicating that almost the entire dataset was successfully included in the topic modeling process.

Similar patterns are observed for the Google Reviews dataset. HDBSCAN excluded around 24.61% of the documents as outliers, while K-Means reduced this exclusion to 4.86%. Meanwhile, BIRCH clustering again showed superior performance, limiting outliers to only 2.64% of documents. This indicates that BIRCH can handle various dataset characteristics effectively, including those with more formal and structured review content. On the YouTube dataset, which contains highly diverse and noisy user-generated comments, HDBSCAN excluded 23.97% of documents as outliers, while K-Means achieved better coverage by reducing outliers to 6.66%. Impressively, BIRCH again demonstrated robust clustering capabilities by reducing outliers to merely 4.84%, indicating strong adaptability even on highly variable and unstructured datasets.

The significantly lower outlier proportions achieved by BIRCH clustering directly contribute to its ability to produce more comprehensive and representative topic models. By ensuring that more documents are included in the clustering process, BIRCH enhances the semantic coverage of the discovered topics, which subsequently improves coherence metrics such as topic diversity, embedding density, and intra-topic similarity as formerly discussed.

Moreover, minimizing outliers also stabilizes the number of topics generated as dataset sizes increase, as reflected in Figure 2, where BIRCH exhibits smoother and more consistent topic count trends. Overall, the outlier handling capability of BIRCH not only ensures better data utilization but also plays a crucial role in enhancing the relevance, consistency, and stability of topic modeling outcomes on Indonesian short text datasets. From Figure 2, BERTopic with BIRCH clustering provides more stable and relevant topic

counts compared with standard BERTopic and BERTopic with K-Means clustering. From Figure 2(a), on X data, BERTopic with BIRCH shows a stable increasing trend and some fluctuations, with the highest peak obtained at 400 texts. Standard BERTopic also shows a peak in the number of topics at 400 texts, then decreases and remains stable at a low number of topics, whereas BERTopic with K-Means shows a low and stable number of topics with the increase in the number of texts. From Figure 2(b), on the Google Review data, BERTopic with BIRCH shows a more stable trend and a more gradual increase in the number of topics, with a peak in the number of topics at around 900 texts. Standard BERTopic exhibits significant fluctuations, which may indicate uncertainty in the topic modeling. BERTopic with K-Means tends to have a relatively low and stable topic count across the range of text counts. From Figure 2(c), on the YouTube data, BERTopic with BIRCH and BERTopic with K-Means show significant fluctuations in the number of topics, which may indicate instability in the topic modeling. Standard BERTopic showed very low performance in generating relevant topics.

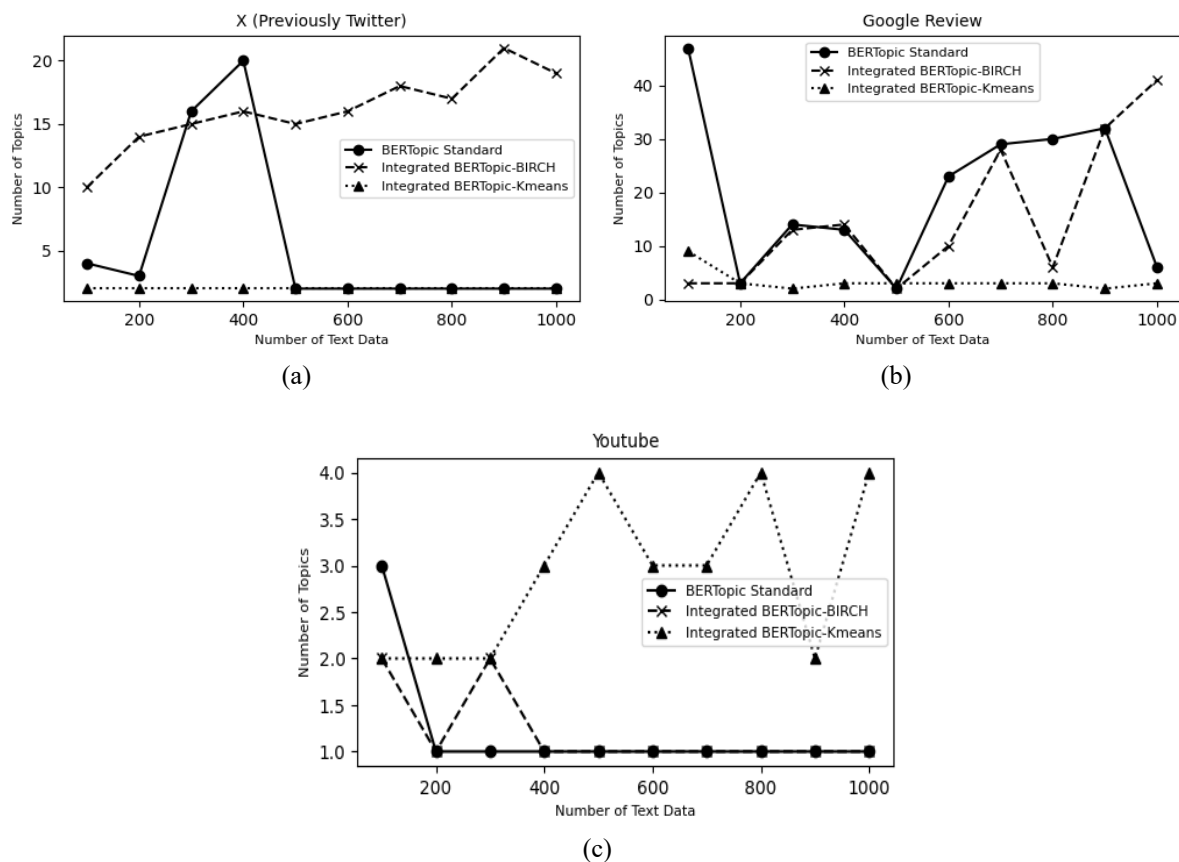


Figure 2. Comparison of the number of topics obtained for datasets from of (a) X, (b) Google Review, and (c) YouTube

Overall, BERTopic with BIRCH generally provides more stable and relevant topic counts on the Google Review and X data, although it exhibits significant fluctuations on the YouTube data. Meanwhile, standard BERTopic shows significant instability on the Google Review and X data and low performance on the YouTube data. BERTopic with K-Means shows a relatively stable but low number of topics, with some fluctuations in the Google Review and YouTube data. The results reveal that improvement is needed to reduce fluctuations in the number of topics generated on the YouTube data.

#### 4. CONCLUSION

This study introduced an enhanced BERTopic framework that integrates IndoBERT embeddings, BM25 weighting, and BIRCH clustering to address key challenges in Indonesian short text topic modeling. The proposed method effectively improves topic coherence while minimizing outlier proportions, leading to

more stable and representative topic structures across diverse datasets. Experimental results demonstrate consistent performance improvements in diversity, embedding density, and intra-topic similarity across multiple data partitions. The framework also maintains robustness and scalability as dataset sizes increase, confirming its suitability for real-world applications involving sparse and dynamic short texts. Future research may extend this framework to model temporal topic evolution or adapt it for multilingual and cross-lingual short text corpora.

ACKNOWLEDGMENTS

Muhammad Muhajir is supported by the Indonesian Education Scholarship (BPI), provided by the Center for Higher Education Funding and Assessment of the Indonesian Ministry of Higher Education, Science, and Technology (PPAT), No. 04018/BPPT/BPI.06/9/2024. His studies benefit from the Indonesia Endowment Fund for Education (LPDP).

FUNDING INFORMATION

Funding information is not available.

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 Muhajir	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓
Gunardi	✓	✓				✓				✓		✓	✓	✓
Danardono	✓		✓	✓			✓			✓			✓	
Dedi Rosadi	✓				✓	✓				✓		✓		

- C : Conceptualization
- M : Methodology
- So : Software
- Va : Validation
- Fo : Formal analysis
- I : Investigation
- R : Resources
- D : Data Curation
- O : Writing - Original Draft
- E : Writing - Review & Editing
- Vi : Visualization
- Su : Supervision
- P : Project administration
- Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest relevant to this paper.

DATA AVAILABILITY

The data supporting this study were collected through web scraping from public sources on X (formerly Twitter), Google Play Store, and YouTube. Due to platform policies and privacy considerations, the data are not publicly available but can be requested from the corresponding author, [MM].

REFERENCES

[1] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *International Journal of Information Technology*, vol. 15, no. 4, pp. 2187–2195, 2023, doi: 10.1007/s41870-023-01268-w.

[2] V. S. Anoop, S. Asharaf, and P. Deepak, "Unsupervised concept hierarchy learning: a topic modeling guided approach," *Procedia Computer Science*, vol. 89, pp. 386–394, 2016, doi: 10.1016/j.procs.2016.06.086.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-AS11>3.0.CO;2-9.

[4] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1–2, pp. 177–196, 2001, doi: 10.1023/A:1007617005950.

[5] A. Goyal and I. Kashyap, "Latent dirichlet allocation - an approach for topic discovery," *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, pp. 97–102, 2022, doi: 10.1109/COM-IT-CON54601.2022.9850912.

[6] L. T. Xu, Z. Xue, and H. Huang, "Short text semantic feature extension and classification based on LDA," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 715, no. 1, doi: 10.1088/1757-899X/715/1/012110.




[7] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: a comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, 2020, doi: 10.3389/frai.2020.00042.






- [8] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 1445–1455, doi: 10.1145/2488388.2488514.
- [9] F. Yi, B. Jiang, and J. Wu, "Topic modeling for short texts via word embedding and document correlation," *IEEE Access*, vol. 8, pp. 30692–30705, 2020, doi: 10.1109/ACCESS.2020.2973207.
- [10] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv-Computer Science*, pp. 1–10, Mar. 2022.
- [11] L. McInnes, J. Healy, and J. Melville, "UMAP: uniform manifold approximation and projection for dimension reduction," *arXiv-Statistics*, Sep. 2020.
- [12] M. Zhou, Y. Kong, and J. Lin, "Financial topic modeling based on the BERT-LDA embedding," in *IEEE International Conference on Industrial Informatics (INDIN)*, 2022, , pp. 495–500, doi: 10.1109/INDIN51773.2022.9976145.
- [13] M. d. Groot, M. Aliannejadi, and M. R. Haas, "Experiments on generalizability of BERTopic on multi-domain short text," *arXiv-Computer Science*, pp. 1–3, Dec. 2022.
- [14] Q. Wang and B. Ma, "Enhancing BERTopic with pre-clustered knowledge: reducing feature sparsity in short text topic modeling," *Journal of Data Analysis and Information Processing*, vol. 12, no. 04, pp. 597–611, 2024, doi: 10.4236/jdaip.2024.124032.
- [15] R. Tomar and A. Sharma, "K-Means and BIRCH: a comparative analysis study," in *Inventive Communication and Computational Technologies*, pp. 281–294, 2023, doi: 10.1007/978-981-19-4960-9\_23.
- [16] F. Jian, J. X. Huang, J. Zhao, Z. Ying, and Y. Wang, "A topic-based term frequency normalization framework to enhance probabilistic information retrieval," *Computational Intelligence*, vol. 36, no. 2, pp. 486–521, 2020, doi: 10.1111/coin.12248.
- [17] X. Shan *et al.*, "GLOW: global weighted self-attention network for web search," in *2021 IEEE International Conference on Big Data, Big Data 2021*, 2021, pp. 519–528, doi: 10.1109/BigData52589.2021.9671546.
- [18] F. Koto, J. H. Lau, and T. Baldwin, "INDOBERTWEET: a pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization," in *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10660–10668, doi: 10.18653/v1/2021.emnlp-main.833.
- [19] Z. Liu, H. Zhang, C. Xiong, Z. Liu, Y. Gu, and X. Li, "Dimension reduction for efficient dense retrieval via conditional autoencoder," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022, pp. 5692–5698, doi: 10.18653/v1/2022.emnlp-main.384.
- [20] S. Sawant, J. Yu, K. Pandya, C. K. Ngan, and R. Bardeli, "An enhanced BERTopic framework and algorithm for improving topic coherence and diversity," in *24th IEEE International Conference on High Performance Computing and Communications, 8th IEEE International Conference on Data Science and Systems, 20th IEEE International Conference on Smart City and 8th IEEE International Conference on Dep*, 2022, pp. 2251–2257, doi: 10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00332.
- [21] N. F. F. d. Silva *et al.*, "Evaluating topic models in Portuguese political comments about bills from Brazil's chamber of deputies," *Intelligent Systems (BRACIS 2021)*, pp. 104–120, 2021, doi: 10.1007/978-3-030-91699-2\_8.
- [22] S. M. Isa, G. Nico, and M. Permana, "IndoBERT for Indonesian fake news detection," *ICIC Express Letters*, vol. 16, no. 3, pp. 289–297, 2022, doi: 10.24507/icicel.16.03.289.
- [23] A. Alzu'Bi and M. Barham, "Automatic BIRCH thresholding with features transformation for hierarchical breast cancer clustering," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1498–1507, 2022, doi: 10.11591/ijece.v12i2.pp1498-1507.
- [24] F. Ramadhani, M. Zarlis, and S. Suwilo, "Improve BIRCH algorithm for big data clustering," *IOP Conference Series: Materials Science and Engineering*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012090.
- [25] A. I. Kadhim, "Term weighting for feature extraction on Twitter: a comparison between BM25 and TF-IDF," *2019 International Conference on Advanced Science and Engineering, ICOASE 2019, Zakho - Duhok, Iraq*, pp. 124–128, 2019, doi: 10.1109/ICOASE.2019.8723825.
- [26] M. Khodorchenko, N. Butakov, and D. Nasonov, "Towards better evaluation of topic model quality," *Conference of Open Innovation Association, FRUCT*, vol. 2022-Novem, pp. 128–134, 2022, doi: 10.23919/FRUCT56874.2022.9953874.
- [27] J. M. Campagnolo, D. Duarte, and G. D. Bianco, "Topic coherence metrics: how sensitive are they?," *Journal of Information and Data Management*, vol. 13, no. 4, 2022, doi: 10.5753/jidm.2022.2181.
- [28] Y. Bu, M. Li, W. Gu, and W. b. Huang, "Topic diversity: a discipline scheme-free diversity measurement for journals," *Journal of the Association for Information Science and Technology*, vol. 72, no. 5, pp. 523–539, 2021, doi: 10.1002/asi.24433.
- [29] I. Rushkin, "Document similarity from vector space densities," *Advances in Intelligent Systems and Computing*, vol. 1251 AISC, pp. 160–171, 2021, doi: 10.1007/978-3-030-55187-2\_14.
- [30] H. Zhao, L. Du, W. Buntine, and M. Zhou, "Supplementary material for 'inter and intra topic structure learning with word embeddings,'" in *35th International Conference on Machine Learning, ICML 2018*, 2018, vol. 13, pp. 9398–9404.
- [31] M. Jesse, C. Bauer, and D. Jannach, "Intra-list similarity and human diversity perceptions of recommendations: the details matter," *User Modeling and User-Adapted Interaction*, vol. 33, no. 4, pp. 769–802, 2023, doi: 10.1007/s11257-022-09351-w.

## BIOGRAPHIES OF AUTHORS






**Muhammad Muhajir**    is currently a doctoral candidate in Mathematics at Universitas Gadjah Mada. He graduated from the Bachelor of Statistics Program, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia in 2011. He earned his Master of Mathematics degree from the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University in August 2014. Until now, he has been working in the Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia. His subjects of specialization include data mining, machine learning, multivariate, data science, and text mining. He can be contacted at email: mmuhajir@uii.ac.id.






**Gunardi**    received his bachelor's degree (Drs.) in Mathematical Statistics and Probability in 1989, his master's degree (M.Si.) in Mathematics with a concentration in Statistics in 1995, and his doctorate (Dr.) in Mathematics with a concentration in Finance in 2007, all from Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia. He is currently a full professor and lecturer at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, and serves as Head of the Statistics Laboratory. Previously, he served as vice dean for finance, asset, and human resources of the faculty from 2016 to 2021. His research interests include applied mathematics, statistics, financial mathematics, data science, machine learning, and mathematical modeling. In addition to his teaching and supervision activities, he has been actively involved in numerous research projects, scientific publications, and academic leadership roles. His professional experience reflects his strong commitment to advancing mathematical applications in academic, financial, and multidisciplinary domains. He can be contacted at email: [gunardi@ugm.ac.id](mailto:gunardi@ugm.ac.id).



**Danardono**    is an Associate Professor in the Department of Mathematics, Universitas Gadjah Mada, Yogyakarta. He graduated from the Statistics Bachelor Program, Faculty of Mathematics and Natural Sciences, Gadjah Mada University (UGM) in 1992. Since 1994, he has worked in the Faculty of Mathematics and Natural Sciences, UGM and is also affiliated with the Clinical Epidemiology Unit, Faculty of Medicine, UGM. He obtained a Master of Public Health in Biostatistics from the Department of Biostatistics and Demography, Faculty of Public Health, Khon Kaen University, Thailand. The field of epidemiology and medicine motivated him to do methodological research in this area for his doctoral study, and in 2005, he received a Ph.D. in Statistics from Umeå University, Sweden. His main research interests are survival data (time-to-event data) analysis and longitudinal data analysis. His research focuses on developing and applying data analysis for epidemiological, medical, and actuarial problems. He can be contacted at email: [danardono@ugm.ac.id](mailto:danardono@ugm.ac.id).



**Dedi Rosadi**    currently works as a (full) professor, leading the Statistical Computing Research Group in the Department of Mathematics, Universitas Gadjah Mada. He graduated from the Statistics Study Program at Gadjah Mada University in February 1996 and started to work as a lecturer at UGM after graduation. In August 1997, he started a Master of Science degree in Stochastic Modeling (Applied Statistics) at the University of Twente, Netherlands, and graduated in June 1999. From September 2001 to September 2004, he started his doctoral study in Econometrics at the Institute of Econometrics and Operation Research (EOS 119), Vienna University of Technology (TU Wien), Austria. After finishing his doctoral study, he came back to Yogyakarta and continued work at the Universitas Gadjah Mada. In September 2013, he received the title of full professor. His subjects of specialization include biostatistics, data science, (statistical) machine learning cluster - statistics, computational statistics, statistical finance, and time series analysis cluster - statistics. He can be contacted at email: [dedirosadi@ugm.ac.id](mailto:dedirosadi@ugm.ac.id).