Scientific Journal of Informatics
Vol. 13, No. 3, August 2026
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Gibbs-BERTopic for Topic Modeling of Short Indonesian Social Media Texts on Artificial Intelligence Issues

Nur'aini[1], Budi Susetyo[2]*, Cici Suhaeni[3]
[1,2,3]Department of Statistics and Data Science, IPB University, Indonesia

Abstract.
Purpose: This study aims to evaluate whether integrating Gibbs Sampling into BERTopic improves topic modeling of short Indonesian social media texts concerning Artificial Intelligence. Baseline BERTopic is used as the primary comparator, while LDA and Top2Vec are included as external baselines representing probabilistic and embedding-based topic-modeling approaches.
Methods: A corpus of 9,585 public posts from platform X, collected from January 2024 to May 2025, was modeled using Latent Dirichlet Allocation (LDA), Top2Vec, Baseline BERTopic, and Gibbs-BERTopic. Model performance was assessed using topic coherence, topic diversity, topic uniqueness, number of topics, and document coverage. The number of outliers was compared specifically between Baseline BERTopic and Gibbs-BERTopic because LDA and Top2Vec do not use an equivalent HDBSCAN-based outlier mechanism.
Result: Baseline BERTopic generated 58 topics and 4,674 outliers, with coherence, diversity, and uniqueness scores of 0.480, 0.460, and 0.380. Gibbs-BERTopic generated 15 topics without outliers and achieved the highest corresponding scores of 0.510, 0.880, and 0.870. LDA and Top2Vec produced lower scores, with 58 and 84 topics, respectively. These findings indicate that, within the corpus and configuration examined, Gibbs-BERTopic provided a more coherent and distinctive topic representation, broader document coverage, and a more compact topic structure.
Novelty: This study extends the evaluation of Gibbs-BERTopic to short Indonesian social media texts and compares it with probabilistic, embedding-based, and BERTopic-based models. The results highlight the importance of evaluating topic models using multiple complementary metrics rather than relying on a single performance measure.

Keywords: Artificial Intelligence, BERTopic, Gibbs-BERTopic, Indonesian Social Media, Topic Modeling

Received May 2026 / **Revised** July 2026 / **Accepted** August 2026

INTRODUCTION

Technological advances, particularly the Internet of Things (IoT), have increased the volume, velocity, and variety of data, which are key characteristics of Big Data [1], [2]. Much of this data is unstructured, including text generated on social media. Text analysis can reveal public opinions, responses, and social dynamics that conventional numerical analysis may not fully capture [3]. Topic modeling is one method for analyzing such data. It identifies latent thematic structures from relationships between words and documents in a corpus [4].

Topic modeling is relevant for analyzing discussions about artificial intelligence in Indonesia. The use of AI in education, healthcare, and industry has generated diverse public opinions, opportunities, and concerns [5], [6]. Social media provides an open forum in which these discussions can be examined systematically. Several methods have been developed to identify thematic structures in text corpora. Advances in transformer-based language representation led to BERTopic. The method combines document embeddings, dimensionality reduction, document clustering, and class-based Term Frequency-Inverse Document Frequency to construct topic representations. Contextual embeddings allow BERTopic to capture semantic similarity between documents rather than relying only on word co-occurrence patterns. This capability is useful for social media texts, where similar ideas may be expressed using different words [7], [8]

[2]*Corresponding author.
Email addresses: nur.aiiniinuraini@apps.ipb.ac.id (Nur'aini), budisu@apps.ipb.ac.id (Susetyo)*, cici_suhaeni@apps.ipb.ac.id (Suhaeni)
DOI: 10.15294/sji.v13i3.50254

However, applying BERTopic to short texts remains challenging. Limited information in short documents can produce sparse representations. As a result, HDBSCAN may form many small clusters or classify documents as outliers. de Groot et al. reported that the proportion of documents classified as outliers could exceed 74% [9]. Wang and Ma also identified feature sparsity as a major challenge in BERTopic-based short-text modeling [10]. Consequently, some documents may be excluded from the final topic structure, and their information may not be represented in the interpretation.

Gibbs Sampling has been used in topic modeling to iteratively update topic assignments based on document-topic and topic-word probability distributions[11]. This procedure has also been integrated into BERTopic to refine the assignments and distributions produced by the initial model. Gibbs-BERTopic has been reported to improve document allocation and produce a more stable topic structure than the initial BERTopic model [12]. However, reducing the number of outliers does not by itself demonstrate better model quality. The result should be interpreted together with within-topic word coherence, across-topic diversity, and the distinctiveness of each topic representation.

Previous applications of Gibbs-BERTopic have been limited to Mandarin-language corpora [12]. These findings cannot be directly generalized to short Indonesian social media texts because linguistic variation and corpus characteristics can affect natural language processing performance [13]. Indonesian also has linguistic features and word-formation patterns that increase lexical variation [14]. These challenges are particularly relevant to social media texts, which are often short, informal, and context-limited.

To address this gap, this study evaluates Gibbs-BERTopic for modeling short Indonesian social media texts about artificial intelligence. Baseline BERTopic is the primary comparator because Gibbs-BERTopic is derived from its framework. LDA and Top2Vec are included as external baselines to represent two additional topic-modeling paradigms. LDA represents a probabilistic approach that models distributions over words and documents [15]. Top2Vec represents a semantic-embedding approach that identifies topics from the proximity of document, word, and topic representations in a vector space [8]. These baselines allow Gibbs-BERTopic to be evaluated not only against its source model but also against lexical and semantic approaches with different topic-formation mechanisms. The models are compared using the number of topics, topic coherence, topic diversity, and topic uniqueness. Document coverage and the number of outliers are compared only between Baseline BERTopic and Gibbs-BERTopic because LDA and Top2Vec do not provide an outlier mechanism directly equivalent to HDBSCAN. This study contributes an evaluation of Gibbs-BERTopic in a new language and corpus setting. It also assesses topic quality using several complementary output characteristics.

METHODS

Data Acquisition and Preprocessing

The data consisted of public posts from the social media platform X. The posts were collected automatically from January 2024 to May 2025. Data collection used Tweet Harvest version 2.7.1, an open-source scraper developed by Satria [16]**.** The software was supported by Playwright and executed in Google Colab. Three search terms were used: "AI," "artificial intelligence," and "kecerdasan buatan." No specific hashtags were used. The search was restricted to Indonesian-language posts and excluded pure retweets. Replies and quote tweets were retained when they contained additional text and at least one search term. The initial collection produced 10,533 posts.

A post was included when it was published during the observation period, written in Indonesian, and contained at least one search term. Relevance screening used word- and phrase-based rules to identify posts that were potentially unrelated to artificial intelligence, contained spam or commercial promotion, or used the abbreviation "AI" in another context. Posts flagged for exclusion were not removed immediately. The first author manually audited their context. Posts that were still related to artificial intelligence were returned to the corpus, whereas confirmed irrelevant posts were excluded. Exact duplicate posts and posts containing only a link were also removed.

The retained documents were preprocessed by removing usernames, symbols, emojis, numbers, and URLs. Case folding and text normalization were then applied. Documents that became empty after preprocessing were excluded. After collection, relevance screening, manual auditing, and preprocessing, 9,585 documents remained in the final modeling corpus.

Latent Dirichlet Allocation (LDA) Baseline
Latent Dirichlet Allocation (LDA) was used as an external baseline representing probabilistic topic modeling based on lexical information. LDA is a generative probabilistic model. It assumes that each document is composed of a mixture of latent topics and that each topic is represented by a probability distribution over the vocabulary. Under this formulation, a document may be associated with several topics in different proportions. The LDA generative process is defined as follows in (1) and (2) [15].

$$\phi_k \sim \text{Dirichlet}\,(\eta),\ \theta_d \sim \text{Dirichlet}\,(\alpha), \tag{1}$$
$$z_{dn} \sim \text{Categorical}\,(\theta_d),\ w_{dn} \sim \text{Categorical}\big(\phi_{z_{dn}}\big) \tag{2}$$

In these equations, ϕ_k denotes the word distribution for topic k, whereas θ_d denotes the topic distribution for document d. The variable z_{dn} is the latent topic assignment for word position n in document d, and w_{dn} is the observed word. The parameters α and η control the priors for the document-topic and topic-word distributions, respectively.

In this study, LDA was applied to the same lexically preprocessed corpus used by the other models. Documents were represented as a document-term matrix based on word frequencies. Several candidate numbers of topics were evaluated. The final configuration was selected using the highest topic Coherence C_V score. Topic diversity and topic uniqueness were then calculated to assess the diversity and distinctiveness of representative words across topics. The LDA configuration and hyperparameters are reported in Table 1.

Top2Vec Baseline
Top2Vec was used as an external baseline representing embedding-based topic modeling. The model represents documents, words, and topics in the same semantic vector space. Proximity between vectors therefore indicates semantic similarity. Topics are identified from dense regions in the document-vector space. The number of dense regions determines the number of topics automatically, and each topic vector is calculated from the centroid of the documents in its cluster [17]. The topic vector for topic k is defined as follows in Equation (3):

$$t_k = \frac{1}{\mid C_k \mid} \sum_{d_i \in C_k} d_i \tag{3}$$

where t_k is the vector for topic k, C_k is the set of documents in cluster k, $\mid C_k \mid$ is the number of documents in that cluster, and d_i is the vector for document i. The most representative words for a topic are identified from the proximity between word vectors and the topic vector. This proximity can be measured using cosine similarity in Equation (4):

$$\text{sim}\big(t_k, w_j\big) = \frac{t_k^T w_j}{\|t_k\|\|w_j\|} \tag{4}$$

where w_j is the vector for word j. Words with the highest cosine similarity are used to represent the topic. This formulation reflects the main principle of Top2Vec: a topic vector is the center of a dense document region, and its topic words are the word vectors closest to that center.

In this study, Top2Vec was applied to the same preprocessed corpus as the other models. Document representations were generated using the multilingual sentence-embedding model paraphrase-multilingual-MiniLM-L12-v2 [18]. UMAP was used to reduce the dimensionality of the document vectors, and HDBSCAN was used for clustering. A grid search evaluated all combinations of the candidate hyperparameters listed in Table 1. The final configuration was selected based on the highest topic coherence score. The number of topics was determined automatically from the resulting document clusters. Topic diversity and topic uniqueness were calculated after the final configuration had been selected.

BERTopic Baseline Model
BERTopic was used as the primary baseline to cluster documents according to semantic similarity [19], [20]. It combines transformer-based representations with density-based clustering. First, the texts are

encoded as high-dimensional vectors using the Siamese Sentence-BERT (S-BERT) architecture. Documents with similar contexts are therefore positioned close to one another in the semantic space [21]. UMAP then reduces the vector dimensions while preserving local structure and latent relationships between documents [22].

The reduced vectors are then clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). This algorithm identifies clusters from semantic density and labels outlying documents as "Topic -1." These outliers provide a basis for comparing the baseline and hybrid models. Finally, class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) is used to extract the most representative keywords from each cluster. The weight of each word in a topic is calculated using Equation (5):

$$\mathrm{W_{t,c}} = \mathrm{tf_{t,c}} \cdot \log\left(1 + \frac{\mathrm{A}}{\mathrm{tf_t}}\right) \tag{5}$$

where $W_{t,c}$ denotes the weight of term t in cluster c, and $tf_{t,c}$ denotes the frequency of term t in that cluster. All documents in a cluster are treated as one combined document. Furthermore, tf_t is the frequency of term t across all clusters, whereas A is the average number of words per class.

Gibbs Sampling Optimization

Gibbs Sampling was used as a probabilistic inference mechanism to update the topic assignment of each word occurrence. It is a Markov Chain Monte Carlo method that iteratively samples from the conditional distribution of each variable while conditioning on the assignments of the other variables [12]. In this study, z_{dn} denotes the topic assignment of word w_{dn}, which occurs at position n in document d.

At each iteration, the previous topic assignment of word w_{dn} is temporarily removed from the document-topic and topic-word counts. The probability of assigning that word to topic k is then calculated in Equation (6) using the following conditional distribution:

$$P(z_{dn'} =, k, \mid Z_{-dn} W \alpha \beta) \propto \frac{n_{d,k}^{-dn} + \alpha}{n_{d,\cdot}^{-dn} + K\alpha} \frac{n_{k,w_{dn}}^{-dn} + \beta}{n_{k,\cdot}^{-dn} + V\beta} \tag{6}$$

The quantity $n_{d,k}^{-dn}$ is the number of other word occurrences in document d assigned to topic k. The quantity $n_{k,w_{dn}}^{-dn}$ is the number of occurrences of word w_{dn} assigned to topic k, excluding the word currently being updated. Furthermore, $n_{d,\cdot}^{-dn}$ is the total number of other topic assignments in document d, and $n_{k,\cdot}^{-dn}$ is the total number of words assigned to topic k. The parameter K is the number of topics, V is the vocabulary size, and α and β are symmetric Dirichlet prior parameters.

The first component of Equation (6) represents the tendency of topic k to occur in document d. The second component represents the tendency of word w_{dn} to occur in topic k. A new topic assignment is sampled from the normalized probability distribution and then added back to the document-topic and topic-word counts. This procedure is repeated for all word occurrences until the specified number of iterations is reached.

After the updating process, the counts from the final assignment state are used to estimate the topic-word distribution as follow in Equation (7):

$$\hat{\phi}_{k,w} = \frac{n_{k,w} + \beta}{n_{k,\cdot} + V\beta} \tag{7}$$

and the document-topic distribution as follow in Equation (8):

$$\hat{\theta}_{d,k} = \frac{n_{d,k} + \alpha}{n_{d,\cdot} + K\alpha} \tag{8}$$

The distribution $\hat{\phi}_{k,w}$ gives the probability of each word within a topic, whereas $\hat{\theta}_{d,k}$ gives the topic proportions for each document. These two distributions are the outputs of Gibbs Sampling and are subsequently used in the Gibbs-BERTopic framework.

Gibbs-BERTopic Framework

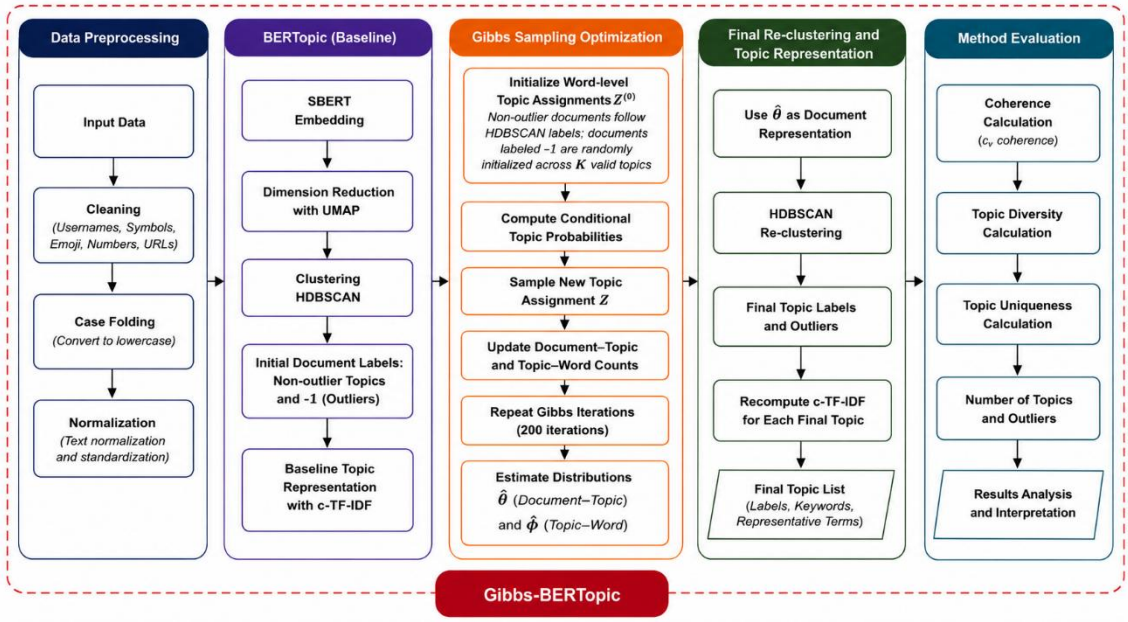

Figure 1 Gibbs-BERTopic framework

Figure 1 show Gibbs-BERTopic framework. Gibbs-BERTopic combines the initial topic structure produced by BERTopic with Gibbs-based topic assignment updates. Baseline BERTopic is first fitted to obtain one initial cluster label for each document. It produces 58 non-outlier topics. These labels are mapped to consecutive topic indices, and Gibbs Sampling is initialized with $K = 58$ valid topics. BERTopic assigns labels at the document level, whereas Gibbs Sampling updates assignments for individual token occurrences. Therefore, each document label is used to initialize token-level topic assignments.

For documents with a valid BERTopic label, each token is initialized to its document topic with probability $1 - \varepsilon$. With probability ε, the token is randomly assigned to one of the K valid topics. This random component allows exploration during Gibbs Sampling. The value of ε, or eps_init_noise, is set to 0.05. Thus, a token follows its document-level topic label with probability 0.95 and receives a random initial topic with probability 0.05. All tokens in documents labeled as HDBSCAN outliers (-1) are randomly initialized to one of the 58 valid topics. Therefore, the initial outlier documents are not excluded from modeling.

All 9,585 documents with at least one token after preprocessing participate in every Gibbs Sampling iteration, including documents labeled -1 during the initial clustering. At each update, the previous topic assignment of a token is removed from the document-topic and topic-word counts. The conditional assignment probability for every topic is then calculated using Equation (4). A new topic is sampled stochastically from the normalized distribution and added back to the count matrices. This procedure is applied to all tokens for 200 iterations, with $\alpha = 0.10$ and $\beta = 0.05$.

After the final iteration, the topic-assignment counts are used to estimate the topic-word distribution $\hat{\phi}$ and the document-topic distribution $\hat{\theta}$ using Equations (5) and (6). Each row of $\hat{\theta}$ represents the topic-probability profile of one document. Each row of $\hat{\phi}$ represents the word probabilities for one topic.

In this implementation, the document-topic distribution matrix $\hat{\theta}$ is used directly as the document representation for the final clustering stage. Unlike the original Gibbs-BERTopic framework, this study does not apply a second UMAP reduction after Gibbs Sampling. HDBSCAN is applied directly to $\hat{\theta}$ to form the final clusters from similarities between document-level topic-distribution profiles. The final number of topics is not specified in advance. It is determined by the density structure of documents in the document-topic space.

This mechanism allows documents initially labeled -1 to receive a new topic label based on their topic-distribution profiles. However, HDBSCAN can still retain a document as an outlier when it does not satisfy the density requirements of any final cluster. Under the configuration used in this study, the final clustering produced 15 topics and no remaining outliers. After the final labels were obtained, c-TF-IDF was recalculated so that the representative words and document labels referred to the same final topic structure.

Model Hyperparameters

Model configurations and hyperparameters were adapted to the characteristics of each method. Parameters with several candidate values were evaluated to identify the final configuration. Parameters marked as fixed were used without tuning. Table 1 reports the configurations of LDA, Top2Vec, Baseline BERTopic, and Gibbs-BERTopic.

Table 1. Model hyperparameters

Model	Parameter	Candidate values	Final value
LDA	ngram_range	(1,1), (1,2)	(1, 2)
	min_df	Fixed	5
	max_df	Fixed	0.90
	Number of Topics ((K))	5, 8, 10, 15, 20, 30, 40, 50, 58, 60	58
	max_iter	20, 50	50
	learning_method	Fixed	batch
Top2Vec	Model embedding	Fixed	paraphrase-multilingual-MiniLM-L12-v2
	min_count	Fixed	5
	topic_merge_delta	0.05; 0.10; 0.15	0.10
	n_neighbors	5, 10, 15, 20	15
	n_components	Fixed	5
	metric	Fixed	cosine
	min_cluster_size	5, 10, 15, 20	15
	min_samples	5, 10, 15, 20	5
	cluster_selection_method	Fixed	eom
	Number of Topics	Automatic	Determined by the model
Baseline BERTopic	ngram_range	(1,1); (1,2)	(1,2)
	Model embedding	Fixed	paraphrase-multilingual-MiniLM-L12-v2
	n_neighbors	5, 10, 15, 50	50
	min_dist	Fixed	0.3
	n_components	Fixed	5
	metric	Fixed	cosine
	min_cluster_size	5, 10, 30	10
	min_samples	5, 20	20
	cluster_selection_method	Fixed	eom
Gibbs-BERTopic	α	0.05; 0.10	0.10
	β	0.03; 0.05; 0.10	0.05
	Number of iterations	100; 200	200
	eps_init_noise	0.03; 0.05; 0.15	0.05

Parameters with several candidate values were evaluated through a grid search over all combinations listed in Table 1. The final configuration for each model was selected using the highest topic coherence score.

Topic diversity and topic uniqueness were not used as selection criteria. They were calculated after the final configuration had been chosen.

Evaluation

All models were applied to the same final corpus of 9,585 preprocessed documents. Topic coherence, topic diversity, and topic uniqueness were calculated using the same evaluation corpus, metric definitions, and number of representative words. The ten highest-ranked words from each topic were used. Model configurations were adapted to the topic-formation mechanism of each method because LDA, Top2Vec, Baseline BERTopic, and Gibbs-BERTopic have different structures and hyperparameters. The number of topics was not fixed across models because the study evaluates the topic structure produced by the best configuration of each model. Outlier counts were compared only between Baseline BERTopic and Gibbs-BERTopic because LDA and Top2Vec do not produce an outlier label directly equivalent to HDBSCAN noise label -1.

The evaluation examined three dimensions of topic quality: internal consistency, global diversity, and the uniqueness of word representations. Topic coherence measures semantic relatedness among words within a topic. Topics composed of closely related words tend to receive higher coherence scores [23]. Topic diversity measures variation in keywords across topics from the proportion of unique words in all topic representations [24]. Topic uniqueness complements this measure by assessing how exclusive the words of a topic are and how rarely they appear in other topics [25]. The number of topics and the outlier proportion were also considered as indicators of compact topic structure and document coverage. Outlier counts were compared only between Baseline BERTopic and Gibbs-BERTopic because LDA and Top2Vec do not provide a directly equivalent HDBSCAN noise label.

RESULT AND DISCUSSION

Exploratory Data Analysis (EDA)

The first exploratory step was to generate a word cloud of the most prominent terms in public discussions about artificial intelligence.

Figure 2. Word cloud of Artificial Intelligence discussions

Figure 2 presents the word cloud generated from the cleaned corpus. Prominent terms include "ai," "chatgpt," "dosen" (lecturer), "anak" (child), "belajar" (learning), "Figure" (image), "gemini," "mahasiswa" (student), and "jurnal" (journal). These terms suggest that public discussion on X extends beyond AI technology in general. It also covers education, academic relationships, generative visual content, and everyday experiences of using AI.

Terms such as "plagiasi" (plagiarism), "parafrase" (paraphrase), "detektor" (detector), and "tulis" (write) indicate concern about academic integrity and changes in learning practices associated with AI. Terms such as "generate," "grok," "video," and "illustrator" also show substantial discussion of generative visual AI. Overall, the exploratory analysis indicates considerable lexical variation and the potential formation of several discourse clusters. Topic modeling is therefore required to identify the thematic structure more systematically. The dataset contains 9,585 documents with an average length of 8.56 words. It can therefore be characterized as a short-text corpus with potential sparsity. A boxplot was used to show the document-length distribution more clearly.

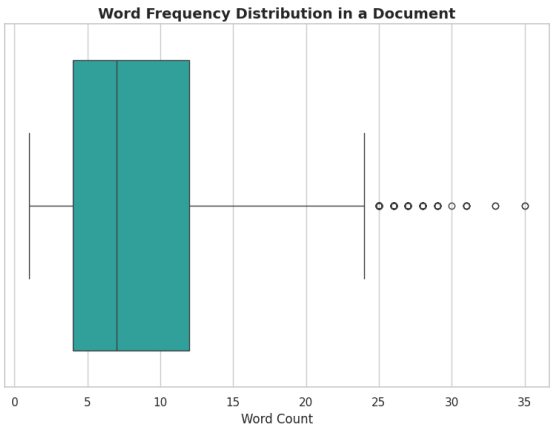

Figure 3. Boxplot number of words in a document

Exploration of the 9,585 preprocessed documents shows that the corpus is short and sparse. As shown in Figure 3, the average document length is approximately 8.6 words, and the median is 7 words. Most documents fall within an interquartile range of 4 to 12 words. The right-skewed distribution indicates that most users express their opinions in highly concise texts, whereas only a small proportion of documents contain up to 35 words. This characteristic limits contextual information and word co-occurrence. It therefore motivates an approach that combines semantic representations with refined topic distributions, such as S-BERT and Gibbs Sampling in the Gibbs-BERTopic framework.

Model Performance Metrics Evaluation

LDA, Top2Vec, Baseline BERTopic, and Gibbs-BERTopic were evaluated using topic coherence, topic diversity, topic uniqueness, and the number of generated topics. Outlier counts were compared only between Baseline BERTopic and Gibbs-BERTopic because LDA and Top2Vec do not use a document-assignment mechanism directly equivalent to HDBSCAN noise detection. Table 2 presents the results.

Table 2. Comparison of evaluation metrics

Metric	LDA	Top2Vec	Baseline BERTopic	Gibbs-BERTopic
Topic Coherence	0.44	0.44	0.48	0.51
Topic Diversity	0.38	0.28	0.46	0.88
Topic Uniqueness	0.27	0.18	0.38	0.87
Number of Topics	58	84	58	15
Number of Outlier Documents	N/A	N/A	4674	0

Table 2 shows that Gibbs-BERTopic achieved the highest topic coherence score at 0.51. Baseline BERTopic followed at 0.48, whereas LDA and Top2Vec each scored 0.44. These results indicate that the representative words within Gibbs-BERTopic topics had relatively stronger internal associations than those produced by the other models. Baseline BERTopic generated 58 topics and classified 4,674 documents as outliers. Thus, almost half of the documents were not included in its non-outlier topics.

Gibbs-BERTopic produced a more compact structure of 15 topics and left no documents as outliers. Its topic coherence increased from 0.48 for Baseline BERTopic to 0.51. The absolute difference was 0.03, which corresponds to a relative increase of 6.25% over the baseline score. This numerical improvement is limited and is therefore interpreted descriptively. Statistical significance could not be assessed because the study did not evaluate variation across repeated runs with multiple random seeds. The score of 0.51 therefore indicates slightly stronger within-topic word associations for the final Gibbs-BERTopic configuration on this corpus. It does not show that the method will always produce higher topic coherence under all conditions.

Topic coherence and topic diversity capture different dimensions of topic quality. Topic coherence measures the relatedness of representative words within each topic. Topic diversity measures the repetition of representative words across all topics. The study in [26] proposed an objective function to improve corpus-level coherence while preserving topic diversity. Their work shows that coherence and diversity should be evaluated jointly rather than treated as interchangeable measures. The present results do not

show a trade-off in which coherence decreases as diversity increases. Gibbs-BERTopic achieved higher topic coherence, topic diversity, and topic uniqueness than Baseline BERTopic. Nevertheless, these metrics should still be interpreted together with the number of topics and document coverage because each describes a different property of the model output.

LDA generated 58 topics with a topic coherence score of 0.44, a topic diversity score of 0.38, and a topic uniqueness score of 0.27. All three scores were lower than those of the two BERTopic variants. Top2Vec generated the largest number of topics, with 84 topics, but obtained the lowest topic diversity and topic uniqueness scores at 0.28 and 0.18, respectively. Under the corpus and configuration examined, the larger number of Top2Vec topics did not result in greater diversity or distinctiveness among representative words.

Topic diversity and topic uniqueness should be interpreted together with the number of topics produced by each model. LDA, Top2Vec, Baseline BERTopic, and Gibbs-BERTopic generated different numbers of topics. Consequently, the number of evaluated representative words and the opportunity for repetition across topics also differed. Topic diversity and topic uniqueness were therefore not used as the sole criteria for model quality. They were interpreted together with topic coherence, the number of topics, and document coverage. The comparison describes model outputs for this corpus and configuration rather than the universal superiority of one method.

Overall, Gibbs-BERTopic produced the most favorable evaluation profile for the corpus and configuration examined. This interpretation considers topic coherence, topic diversity, topic uniqueness, the number of topics, and document coverage together. The model achieved the highest scores on all three topic-quality metrics. It also produced a more compact topic structure and included all documents. Under this configuration, the Gibbs-BERTopic pipeline was associated with broader document coverage and topic representations that were more coherent, diverse, and distinctive. However, these findings are specific to the corpus, embedding model, and configuration used in this study. They should not be generalized as universal superiority of Gibbs-BERTopic.

Visual Comparison of Baseline BERTopic and Gibbs-BERTopic Topic Structures

The topic structures produced by Baseline BERTopic and Gibbs-BERTopic were compared visually using inter-topic distance maps. These visualizations provide supporting evidence about the relative proximity and separation of topics in the two focal models. In BERTopic visualizations, topic representations based on c-TF-IDF are projected into two dimensions using UMAP [22].

LDA and Top2Vec were not included in this visual comparison because they represent topics differently. LDA represents a topic as a probability distribution over the vocabulary [15]. Top2Vec represents a topic as a vector in a shared semantic-embedding space with documents and words [17]. Therefore, distances and overlaps in their visualizations are not directly comparable with BERTopic inter-topic distance maps. A cross-model visual comparison would require standardized topic representations, distance measures, and projection methods.

Figure 4 shows that Baseline BERTopic generated 58 topics. Several topic bubbles are close to one another or overlap in the two-dimensional space. This pattern indicates relative proximity between topics in the projected representation. However, overlap on the map does not by itself demonstrate semantic redundancy because the topic positions are also affected by dimensionality reduction.

Figure 5 shows the more compact Gibbs-BERTopic structure of 15 topics. Several topics appear to be more clearly separated than those produced by Baseline BERTopic. This visual pattern is descriptively consistent with the increase in topic diversity from 0.46 to 0.88 and in topic uniqueness from 0.38 to 0.87. The visualization supports the interpretation, but model quality is assessed from the full set of evaluation metrics.

Figure 4. Baseline BERTopic intertopic distance map

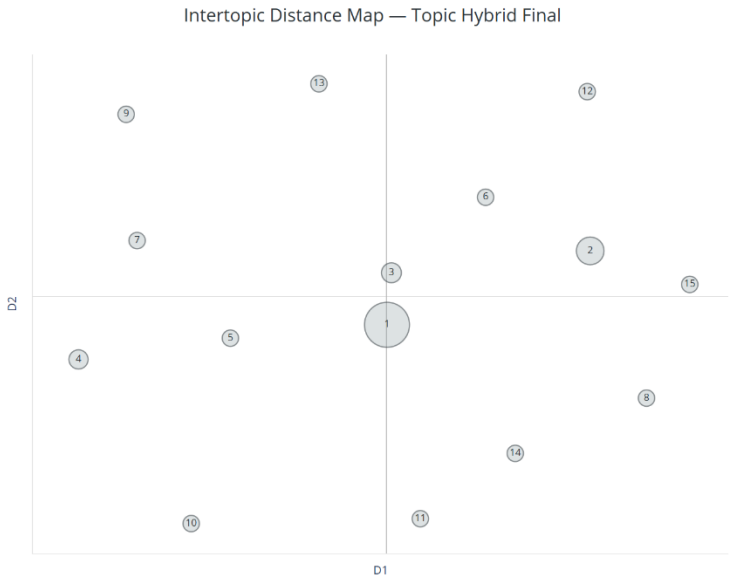

Figure 5. Gibbs-BERTopic intertopic distance map

Methodological Implications for Analyzing Indonesian Discourse

The findings show that topic modeling of short Indonesian social media texts should be evaluated using several output characteristics. Social media corpora are often short, informal, and context-limited. These properties can make it difficult to construct dense document representations. In BERTopic, this limitation may affect density-based clustering and cause documents to be classified as outliers [9], [10]. In this study, Baseline BERTopic generated 58 topics with a topic coherence score of 0.48 and classified 4,674 documents as outliers. Thus, almost half of the corpus was not represented in its non-outlier topics.

Gibbs-BERTopic produced a different output profile. It generated a more compact structure of 15 topics and left no documents as outliers under the selected configuration. Compared with Baseline BERTopic,

topic coherence increased from 0.48 to 0.51, topic diversity increased from 0.46 to 0.88, and topic uniqueness increased from 0.38 to 0.87. Descriptively, these results indicate stronger within-topic word associations, greater overall diversity among representative words, and more distinctive word representations for individual topics. The smaller number of topics also indicates a more compact structure without reducing document coverage.

The external baselines provide additional context. LDA generated 58 topics with topic coherence, topic diversity, and topic uniqueness scores of 0.44, 0.38, and 0.27, respectively. Top2Vec generated the largest number of topics, with 84 topics, and obtained scores of 0.44, 0.28, and 0.18. Under the corpus and configurations examined, neither a probabilistic approach nor an embedding-based approach automatically produced a compact, diverse, and nonredundant topic structure. Output quality also depended on the mechanisms used for document representation, topic formation, and document assignment.

Methodologically, the integration of Gibbs Sampling allows all documents to participate in updates of the document-topic and topic-word distributions, including those labeled -1 during the initial clustering [12]. Initial outliers can therefore acquire new topic representations before HDBSCAN is applied again. In this study, the full pipeline reduced the number of topics from 58 to 15, removed all remaining outliers, and increased all three topic-quality metrics. However, the study did not isolate the contributions of Gibbs Sampling, document-topic distribution representations, HDBSCAN reclustering, and c-TF-IDF recalculation. Therefore, the changes in performance cannot be attributed causally to one modeling stage alone.

The linguistic characteristics of Indonesian are also important when interpreting the results. Affixation, reduplication, variation in word forms, and informal language on social media can increase lexical and semantic variation in a corpus [13], [14]. However, the study design did not separate language effects from the effects of document length, preprocessing, the embedding model, or the clustering configuration. The results therefore do not show that Gibbs-BERTopic is specifically more adaptive to Indonesian morphology. They provide empirical evidence that the method produced a more compact topic structure, included all documents, and achieved higher evaluation scores on the corpus examined.

This study evaluates Gibbs-BERTopic on short Indonesian social media texts, with Baseline BERTopic as the primary comparator and LDA and Top2Vec as external baselines. Gibbs-BERTopic achieved the highest topic coherence, topic diversity, and topic uniqueness scores. It also produced fewer topics and broader document coverage. Nevertheless, topic-model quality should be interpreted from several indicators, including coherence, diversity, distinctiveness, the number of topics, and document coverage. Generalization should remain cautious because the study covers one issue, one social media platform, one embedding model, and a specific set of model configurations. Future studies should evaluate stability across repeated runs, alternative hyperparameters, different embedding models, ablation analyses, and corpora from multiple domains and platforms.

Topic Keyword Visualization

Figure 6 presents the ten words with the highest c-TF-IDF scores for each Gibbs-BERTopic topic. Bar length indicates how strongly a word distinguishes a topic from the other topics. It does not represent the word's frequency in the full corpus. These representative words provide an initial basis for topic interpretation and labeling. The 15 topics were then grouped interpretively into four dimensions based on substantive similarity. This grouping was used only to facilitate presentation and was not an additional clustering output.

Figure 6 presents the ten words with the highest c-TF-IDF scores for each Gibbs-BERTopic topic. One researcher manually assigned labels to the 15 topics using these words and representative documents from each topic. The same researcher then grouped the topics into four dimensions based on substantive similarity. No additional annotators participated, so inter-rater agreement was not calculated. Table 3 reports the topic labels and dimensions.

Figure 6. Final topic visualization

Table 3 Topic labels

Topic	Topic label	Dimension
1	AI as a general social phenomenon	User relationships with and competence in AI
2	User experiences with chatbots	User relationships with and competence in AI
3	AI in healthcare	Sectoral applications
4	Critiques of AI-generated visuals	Impact of AI on the creative sector
5	AI and the creative industry	Impact of AI on the creative sector
6	AI as an information source	Sectoral applications
7	AI-generated image manipulation	Impact of AI on the creative sector
8	AI in children's education	Sectoral applications
9	AI television presenters	Impact of AI on the creative sector
10	AI in design and visual communication education	Sectoral applications
11	AI infrastructure	AI infrastructure and technical aspects
12	Prompt engineering	AI infrastructure and technical aspects
13	VTubers and virtual idols	Impact of AI on the creative sector
14	AI output quality	Impact of AI on the creative sector
15	AI for work	Sectoral applications

Table 3 organizes the Gibbs-BERTopic output into four main dimensions: user relationships with and competence in AI, sectoral applications, the impact of AI on the creative sector, and AI infrastructure and technical aspects. This grouping supports substantive interpretation and is not an additional model-generated clustering result.

The dimension of user relationships with and competence in AI includes Topics 1 and 2. Topic 1 represents AI as a general social phenomenon discussed through users responses and evaluations of AI technology. Topic 2 focuses on users experiences with chatbots, both as conversational companions and as tools for processing information and text. Together, these topics show that AI is discussed not only as a technology but also as part of everyday user experience.

The sectoral-applications dimension includes Topics 3, 6, 8, 10, and 15. These topics cover the use of AI in healthcare, information retrieval, children's education, design and visual communication education, and work. The dimension shows that AI use in the corpus is associated with practical needs, including obtaining answers, supporting learning, and completing tasks and work activities.

The impact of AI on the creative sector dimension includes Topics 4, 5, 7, 9, 13, and 14. These topics cover criticism of AI-generated visuals, the influence of AI on creative industries, image manipulation and authenticity, AI television presenters, VTubers and virtual idols, and evaluations of AI output quality. The structure indicates that discussion of AI in creative fields concerns more than technical capability. It also addresses content authenticity, changing forms of entertainment, and effects on creative work.

The AI infrastructure and technical aspects dimension includes Topics 11 and 12. Topic 11 concerns supporting infrastructure, including data centers, servers, cooling systems, and resource requirements. Topic 12 concerns prompt engineering as the skill of formulating instructions to obtain a desired model output. These topics show that the corpus also contains discussion of the physical and technical foundations required to operate AI systems.

CONCLUSION

The evaluation of LDA, Top2Vec, Baseline BERTopic, and Gibbs-BERTopic on 9,585 Indonesian social media documents shows that the models produced different topic structures. Baseline BERTopic achieved a topic coherence score of 0.48, generated 58 topics, and classified 4,674 documents as outliers. Gibbs-BERTopic produced a more compact structure of 15 topics and assigned all documents to final topics. It achieved topic coherence, topic diversity, and topic uniqueness scores of 0.51, 0.88, and 0.87, respectively. These values were higher than the Baseline BERTopic scores of 0.48, 0.46, and 0.38. Under the corpus and configuration examined, Gibbs-BERTopic therefore produced topics that were more coherent, diverse, and distinctive, with broader document coverage.

The external baselines provide further context. LDA generated 58 topics with topic coherence, topic diversity, and topic uniqueness scores of 0.44, 0.38, and 0.27, respectively. Top2Vec generated 84 topics with corresponding scores of 0.44, 0.28, and 0.18. Under the corpus and configurations examined, Gibbs-BERTopic provided a better balance between compact topic structure, document coverage, diversity among representative words, and topic distinctiveness. The integration of Gibbs Sampling therefore shows potential for addressing document-allocation limitations in Baseline BERTopic, particularly for documents initially labeled as outliers.

The interpretation of the 15 Gibbs-BERTopic topics identified four main dimensions of AI discussion: user relationships with and competence in AI, sectoral applications of AI, the impact of AI on the creative sector, and AI infrastructure and technical aspects. The topics included user experiences with chatbots, AI as an information source, applications in education and work, the quality and authenticity of visual content, changes in creative industries and virtual entertainment, data-center infrastructure, and prompt engineering. These findings represent documents collected using the study's keywords, platform, and observation period. They are not intended to describe the entire AI discourse in Indonesia.

The findings should be generalized cautiously because the analysis is limited to one issue, one social media platform, one embedding model, and a specific set of configurations and hyperparameters. Topic labeling and grouping were also conducted manually by one researcher and therefore remain subjective. Future research should assess stability across repeated runs and alternative hyperparameters. It should also involve multiple annotators, compare different embedding models, and evaluate corpora from multiple issues, platforms, and varieties of Indonesian text.

REFERENCES

[1] P. Hikmah Febryan, A. Kusuma Negara, and M. Farell Altivan Ramadhan, "ANALISIS PENGGUNAAN AI DALAM ALGORITMA SOSIAL MEDIA : SYSTEMATIC LITERATURE REVIEW," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, pp. 1095–1102, Dec. 2024, doi: 10.36040/jati.v9i1.12613.

[2] A. R. Al-Ali, R. Gupta, I. Zualkernan, and S. K. Das, "Role of IoT technologies in big data management systems: A review and Smart Grid case study," *Pervasive Mob. Comput.*, vol. 100, p. 101905, May 2024, doi: 10.1016/j.pmcj.2024.101905.

[3] A. W. Octavianto, A. Priyonggo, and Y. P. Setianto, "Framing The Future: Exploring AI Narratives in Indonesian Online Media Using Topic Modelling," *Jurnal Komunikasi Indonesia*, vol. 13, no. 2, pp. 172–194, Aug. 2024, doi: 10.7454/jkmi.v13i2.1245.

[4] D. Choirinnisa *et al.*, "LDA Topic Modeling: Twitter-Based Public Opinion on Indonesian Ministry of Finance," *Sinkron*, vol. 9, no. 2, pp. 849–863, May 2025, doi: 10.33395/sinkron.v9i2.14719.

[5] Y. S. Pongtambing *et al.*, "Peluang dan Tantangan Kecerdasan Buatan Bagi Generasi Muda," *Bakti Sekawan : Jurnal Pengabdian Masyarakat*, vol. 3, no. 1, pp. 23–28, Jun. 2023, doi: 10.35746/bakwan.v3i1.362.

[6] B. Herdian Nugroho, "Analisis Komparatif Kemampuan Kecerdasan Buatan dalam Menganalisis Konten Media Sosial: Studi Kasus Grok (XAI), ChatGPT, dan Gemini," *LogicLink*, pp. 56–69, Jun. 2025, doi: 10.28918/logiclink.v2i1.10819.

[7] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar. 2022.
[8] L. Ma *et al.*, “AI-powered topic modeling: comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women,” *Exp. Biol. Med.*, vol. 250, Feb. 2025, doi: 10.3389/ebm.2025.10389.
[9] E. Zve, B. Icard, A. Breton, L. Sainero, G. Bourgne, and J.-G. Ganascia, “From outliers to topics in language models: Anticipating trends in news corpora,” in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, 2025, pp. 385–398.
[10] 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.
[11] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl_1, pp. 5228–5235, Apr. 2004, doi: 10.1073/pnas.0307752101.
[12] Y. Zhu and Y. Liu, “Gibbs-BERTopic: A Hybrid Approach for Short Text Topic Modeling,” *IEEE Access*, vol. 13, pp. 49162–49173, 2025, doi: 10.1109/ACCESS.2025.3552221.
[13] R. Hasnul Ulya, “Transformasi Makrolinguistik Bahasa Indonesia dalam Gamitan Media Digital: Analisis Wacana Kritis pada Platform Media Sosial,” *Jurnal Ilmiah Langue and Parole*, vol. 8, no. 1, pp. 91–99, Dec. 2024, doi: 10.36057/jilp.v8i1.717.
[14] A. S., T. Eugene, Misara, and E. Nurhayati, “Tantangan Linguistik dalam Pengimplementasian Big Data Berbahasa Indonesia pada Robot Humanoid: Tinjauan dan Rekomendasi,” *Jurnal Pendidikan Bahasa dan Sastra Indonesia*, vol. 1, no. 1, p. 9, Dec. 2024, doi: 10.47134/jpbsi.v1i1.1278.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
[16] Helmi Satria, “Tweet Harvest: Twitter crawler.” Accessed: Aug. 07, 2026. [Online]. Available: helmisatria/tweet-harvest
[17] D. Angelov, “Top2Vec: Distributed Representations of Topics,” Aug. 2020.
[18] N. Reimers and I. Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 4512–4525. doi: 10.18653/v1/2020.emnlp-main.365.
[19] N. C. Hellwig, J. Fehle, M. Bink, T. Schmidt, and C. Wolff, “Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election,” *Int. J. Speech Technol.*, vol. 27, no. 4, pp. 901–921, Dec. 2024, doi: 10.1007/s10772-024-10142-4.
[20] D. Medvecki, B. Bašaragin, A. Ljajić, and N. Milošević, “Multilingual Transformer and BERTopic for Short Text Topic Modeling: The Case of Serbian,” 2024, pp. 161–173. doi: 10.1007/978-3-031-50755-7_16.
[21] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.
[22] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Sep. 2020.
[23] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.
[24] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, “Topic Modeling in Embedding Spaces,” *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 439–453, Dec. 2020, doi: 10.1162/tacl_a_00325.
[25] F. Nan, R. Ding, R. Nallapati, and B. Xiang, “Topic Modeling with Wasserstein Autoencoders,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 6345–6381. doi: 10.18653/v1/P19-1640.
[26] R. Li, F. Gonzalez-Pizarro, L. Xing, G. Murray, and G. Carenini, “Diversity-Aware Coherence Loss for Improving Neural Topic Models,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 1710–1722. doi: 10.18653/v1/2023.acl-short.145.