

Joint Optimization of Preprocessing and HDBSCAN Hyperparameters in BERTopic for Indonesian Social Media Text

Sabrina Adnin Kamila¹, Utami Dyah Syafitri^{1*}, Cici Suhaeni¹

¹Department of Statistics and Data Science, IPB University, Indonesia

utamids@apps.ipb.ac.id

ABSTRACT

Article History:

Received : 03-07-2026

Revised : 22-07-2026

Accepted : 03-09-2026

Online : 22-10-2026

Keywords:

BERTopic;

HDBSCAN;

Hyperparameter

Optimization;

Topic Modeling;

Social Media Analytics.



Public discourse on social media offers higher education institutions a large-scale, real-time source of feedback, yet extracting interpretable themes from short and informal Indonesian text remains methodologically challenging. This study proposed a statistically grounded framework for HDBSCAN hyperparameter optimization in BERTopic topic modeling, applied to Indonesian-language social media discourse about IPB University on X. This study aimed to determine how preprocessing decisions and HDBSCAN hyperparameters jointly shape topic quality, and to benchmark the resulting optimized BERTopic model against a default configuration and an LDA baseline. A total of 21,298 posts collected between August 2023 and August 2025 were processed under three preprocessing scenarios. For each scenario, a grid search over 126 HDBSCAN hyperparameter combinations was evaluated using topic coherence, topic uniqueness, document coverage, and factuality, standardized and analyzed with Principal Component Analysis and qualitative inspection of the intertopic distance map. The optimized models were benchmarked against a default-configuration BERTopic model and a Latent Dirichlet Allocation (LDA) baseline. Results showed that default HDBSCAN settings produced 334 to 367 statistically unstable topics per scenario, whereas optimization increased coherence by 40% to 66% and reduced the topic count to three to six interpretable themes. The excess-of-mass (*eom*) method consistently outperformed *leaf*, and the optimal minimum cluster size shifted from 300 to 500 as preprocessing intensity increased, reflecting a direct interaction between preprocessing and clustering. Optimized BERTopic achieved higher coherence than LDA (0.522–0.634 versus 0.506–0.528), while LDA achieved higher topic uniqueness and coverage, a trade-off linked to BERTopic's outlier-labeling mechanism. Dominant themes across both models concerned admission and campus life, tuition, and campus facilities, with LDA also surfacing IPB University's agricultural identity. By treating preprocessing and clustering hyperparameters as a jointly optimized rather than separately tuned design space, these findings offer a replicable framework for topic modeling of short, informal Indonesian text and practical insight for social media analytics and data-driven reputation management.



Crossref

<https://doi.org/10.31764/jtam.v10i4.40667>



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

A. INTRODUCTION

Social media platforms, particularly X (formerly Twitter), have become a central arena for public discourse, where users freely express opinions on institutions, policies, and social issues (Chauhan & Shah, 2022). For higher education institutions, this stream of user-generated content offers unsolicited, real-time feedback on academic reputation and campus experience

at a scale rarely captured through conventional evaluation (Romero et al., 2024). Universities increasingly treat social media as a strategic channel for institutional communication and reputation management (Capriotti & Zeler, 2023), and mining platforms such as X can surface public sentiment and dominant concerns about educational services that traditional surveys capture only partially (Brandon, 2023). However, the volume and unstructured nature of such data render manual analysis impractical, motivating computational approaches that systematically extract latent thematic patterns from large text collections.

Topic modeling is one of the most widely adopted statistical approaches for this purpose, and comprehensive surveys have documented its rapid methodological development and broad application across domains (Abdelrazek et al., 2023; Vayansky & Kumar, 2020). Latent Dirichlet Allocation (LDA) is a generative probabilistic model that represents each document as a mixture of latent topics drawn from a Dirichlet distribution and each topic as a distribution over words, inferring topic structure from word co-occurrence through variational inference or Gibbs sampling (Blei et al., 2003). Owing to its interpretable probabilistic formulation, LDA remains extensively applied, including for Indonesian-language public discourse such as commentary on Indonesia's free nutritious meals program (Suhaeni et al., 2025).

Despite this statistical maturity, LDA's reliance on word co-occurrence limits its performance on short and informal text, where sparse context reduces the reliability of co-occurrence-based estimation and produces overlapping, difficult-to-interpret topics (Albalawi et al., 2020; Weisser et al., 2023). A systematic review of topic models for short-text social media analysis similarly identifies data sparsity as a recurring obstacle to interpretability (Laureate et al., 2023; Zimmermann et al., 2024). Further demonstrated that this limitation persists even when preprocessing is carefully refined, indicating that it originates from LDA's underlying statistical assumptions rather than from data preparation alone.

These limitations have motivated transformer-based alternatives, most notably BERTopic (Grootendorst, 2022). Rather than relying on co-occurrence counts, BERTopic represents each document with dense contextual embeddings produced by pre-trained models such as BERT, which encode bidirectional semantic context across a sentence (Devlin et al., 2019). These embeddings are reduced in dimensionality and clustered using a density-based algorithm, after which each cluster is characterized through a class-based term frequency-inverse document frequency (c-TF-IDF) procedure that adapts TF-IDF weighting from the document level to the cluster level, emphasizing terms statistically distinctive to a topic while down-weighting terms occurring uniformly across topics (Grootendorst, 2022). A complete formulation, together with the clustering and evaluation metrics used in this study, is presented in the Methods section.

By grouping documents according to semantic proximity rather than surface-level vocabulary, BERTopic has been shown to consistently produce more coherent and interpretable topics than LDA and other classical methods on short, informal text (Egger & Yu, 2022; Ma et al., 2025; Babalola et al., 2024). This advantage extends beyond English-language corpora, with comparable improvements reported for short-text data in Hindi (Mutsaddi et al., 2025) and Serbian (Medvecki et al., 2024).

Nevertheless, BERTopic's performance is highly sensitive to the configuration of its density-based clustering component, HDBSCAN. Medvecki et al. (2024) reported that default settings can leave more than 70% of documents unassigned as outliers, while Borčin & Jose (2024)

showed that default configurations are not robust and can substantially reduce topic coherence. Janssens et al. (2025) analyzed the effect of HDBSCAN parameters and topic-reduction techniques on topic count, outlier ratio, coherence, and diversity, while Feng et al. (2025) proposed a targeted optimization scheme that yielded measurable coherence gains. That scheme, however, fixed a single preprocessing pipeline and was validated on one health-communication corpus, so it leaves open whether the gains hold once preprocessing is allowed to vary and whether they transfer to a different language and discourse domain.

A further gap concerns the statistical interaction between preprocessing and clustering. Zimmermann et al. (2024) showed that preprocessing choices alter document representation and, in turn, topic structure, while Farea et al. (2024) found that the optimal number of topics depends jointly on model configuration and the evaluation metric applied. This is particularly relevant for Indonesian social media text, which exhibits code-switching between Indonesian and English, phonetic spelling variation, informal abbreviations, and community-specific terms that standard preprocessing pipelines are not designed to handle (Suhaeni et al., 2025b). Because HDBSCAN operates directly on the embedding geometry that preprocessing produces, tuning the two independently risks a configuration optimal for one pipeline yet unstable for another, which is why preprocessing and clustering hyperparameters are examined here as a single coupled design space.

To address these gaps, this study proposes a systematic optimization framework for BERTopic that jointly evaluates three preprocessing scenarios against a comprehensive grid search over HDBSCAN hyperparameter combinations, with model selection guided by a multi-criteria protocol integrating topic coherence, topic uniqueness, document coverage, and factuality, supplemented by qualitative inspection of the intertopic distance map. The framework is applied to public discourse concerning the reputation of IPB University on X, collected using the keywords "IPB," "ipebe," and "aipibi" between August 2023 and August 2025. Its novelty lies in optimizing preprocessing and HDBSCAN configuration jointly rather than separately, and in providing a replicable multi-criteria framework for short, informal Indonesian institutional discourse, a setting not addressed by prior BERTopic optimization work that fixed preprocessing or targeted a single domain (Feng et al. 2025; Janssens et al. 2025). Accordingly, this study aims to examine how preprocessing decisions and HDBSCAN hyperparameters statistically interact to shape topic quality, to benchmark the optimized model against a default-configuration BERTopic model and an LDA baseline across quantitative and qualitative criteria, and to map the dominant themes in this discourse to support data-driven reputation management.

B. METHODS

This study proposes a systematic optimization framework for BERTopic applied to Indonesian-language social media discourse concerning IPB University, combining three preprocessing scenarios with a grid search over HDBSCAN hyperparameters, with model selection guided by a multi-criteria statistical protocol and the optimized configuration benchmarked against a default-configuration BERTopic model and an LDA baseline. All experiments were implemented in Python 3 and executed on a GPU.

1. Data Collection

Data were collected from X (formerly Twitter) using tweet-harvest, a command-line tool for keyword-based retrieval of public posts. The dataset comprises 21,298 Indonesian-language posts referencing IPB University, collected with the keywords “IPB,” “ipebe,” and “aipibi” over the period from August 2023 to August 2025. Retweets, duplicates, and posts unrelated to IPB University were removed so that each remaining record reflects an independent expression of public opinion.

2. Text Preprocessing

Three preprocessing scenarios were applied independently to statistically evaluate how different text transformation strategies affect the resulting topic structure. Each scenario applied the same four core steps: (1) *cleaning*, comprising HTML-tag and entity removal, case folding, and removal of URLs, mentions, hashtags, punctuation, and non-ASCII characters; (2) *normalization*, comprising slang normalization and word-elongation correction using indoNLP, supplemented by a manually curated dictionary for abbreviations and misspellings not covered by the standard slang lexicon; (3) *stopword removal* using indoNLP, supplemented with a manually compiled list of informal particles and social-media filler words; and (4) *stemming* using PySastrawi. Across all three steps, a protected list of IPB University-specific abbreviations (faculties, study programs, admission pathways) was excluded from transformation to prevent semantic distortion. This choice of rule-based tools was informed by empirical evidence that rule-based methods outperform large language models on stopwords-removal and stemming tasks for Indonesian social media text (Suhaeni et al., 2025b). The three scenarios differed only in how numerical tokens, single-character tokens, and compound expressions were treated, as summarized in Table 1.

Table 1. Text Preprocessing Steps Applied Across Three Scenarios

Preprocessing Step	Scenario 1	Scenario 2	Scenario 3
Noise removal, case folding, slang normalization, stopwords removal, stemming	Applied	Applied	Applied
Numerical tokens	Retained	Removed	Merged with context
Single-character tokens	Retained	Removed	Retained
Word merging (compound terms)	Not applied	Not applied	Applied

Scenario 1 served as the baseline pipeline and retained numerical and single-character tokens. Scenario 2 removed these to reduce vocabulary sparsity and noise. Scenario 3 additionally merged frequently co-occurring compound expressions into single tokens. Preprocessing was treated as part of the model’s design space, rather than a routine cleaning step, following evidence that preprocessing choices statistically alter document representation and, in turn, the resulting topic structure (Zimmermann et al., 2024).

3. LDA Baseline

LDA was implemented as a baseline using the gensim library, under the same three preprocessing scenarios to ensure comparability. LDA represents each document as a mixture of K latent topics, each modeled as multinomial word distribution $\phi_k \sim \text{Dirichlet}(\beta)$. For each document d , a topic-proportion vector $\theta_d \sim \text{Dirichlet}(\alpha)$ is drawn, from which each word's topic assignment $z_{d,n}$ and observed word $w_{d,n}$ are generated in turn. The resulting joint distribution is (Blei et al., 2003):

$$p(\theta, \phi, z, w \mid \alpha, \beta) = \prod_{k=1}^K p(\phi_k \mid \beta) \prod_{d=1}^M p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \phi_{z_{d,n}}) \quad (1)$$

with $d = 1, \dots, M$ (documents), $k = 1, \dots, K$ (topics), and n indexing word positions within document d .

Parameters were estimated using variational inference as implemented in gensim, with the Dirichlet hyperparameters α and β set to “auto” so that both were estimated directly from the data, and training parameters *passes* = 10 and *iterations* = 100 held constant across all runs. For each preprocessing scenario, the number of topics K was varied from 1 to 10, and the model achieving the highest topic coherence score (Equation 8) was selected as the final LDA model.

4. BERTopic Framework

BERTopic combines four sequential components, namely sentence embedding, dimensionality reduction, density-based clustering, and class-based term weighting, to statistically group documents by semantic similarity rather than surface-level vocabulary (Grootendorst, 2022).

a. Sentence Embedding

Each document was transformed into a fixed-length vector using Sentence-BERT (SBERT), which fine-tunes BERT within a siamese network so that embeddings can be compared via cosine similarity (Reimers & Gurevych, 2019). The multilingual variant *paraphrase-multilingual-MiniLM-L12-v2* mapped each document to a 384-dimensional vector across more than 50 languages.

b. Dimensionality Reduction

The resulting high-dimensional embeddings were reduced using Uniform Manifold Approximation and Projection (UMAP), which builds a fuzzy topological representation of the data and optimizes a low-dimensional layout that preserves it (McInnes et al., 2020). For each point x_i , the probabilistic connectivity to a neighbor x_j is computed as:

$$p_{j|i} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \quad (2)$$

where ρ_i is the local connectivity threshold, σ_i is a scale parameter keeping entropy constant, and $d(x_i, x_j)$ is the distance between the two points. In the low-dimensional

embedding, similarity between projected points y_i and y_j is computed with a Cauchy-type kernel:

$$q_{ij} = (1 + a \cdot ||y_i - y_j||^{2b})^{-1} \quad (3)$$

with a and b as constants shaping the curve of the similarity decay. UMAP then minimizes the divergence between p_{ij} and q_{ij} through a binary cross-entropy objective:

$$\mathcal{L} = \sum_{(i,j)} \left[p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right) + (1 - p_{ij}) \log \left(\frac{1 - p_{ij}}{1 - q_{ij}} \right) \right] \quad (4)$$

optimized via stochastic gradient descent, so that points that are close in the original space remain close after reduction (McInnes et al., 2020). The reduction was configured with $n_neighbors = 30$, $n_components = 5$, $min_dist = 0.0$, the *cosine* distance metric, and a fixed *random_state* of 42 to make the stochastic embedding reproducible.

c. Clustering

The reduced embeddings were clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which does not require the number of clusters to be specified in advance (McInnes et al., 2017). HDBSCAN first computes the *core distance* of each point p , defined as its distance to its k -th nearest neighbor, with k set equal to *min_samples*. From this, the *mutual reachability distance* between two points p and q is defined to avoid bias from inconsistent local density:

$$mreach(p, q) = \max\{coredist(p), coredist(q), dist(p, q)\} \quad (5)$$

The mutual-reachability graph builds a minimum spanning tree, from which candidate clusters are derived and selected by a stability score:

$$Stability(C) = \sum_{p \in C} (\lambda_{max}(p) - \lambda_{birth}(C)) \quad (6)$$

where $\lambda = 1/distance$ denotes the density level, $\lambda_{birth}(C)$ is the density level at which cluster C first forms, and $\lambda_{max}(p)$ is the maximum density level at which point p still belongs to C . Clusters with the highest stability are retained, allowing HDBSCAN to determine the number of clusters automatically while making its output highly sensitive to hyperparameter configuration (McInnes et al., 2017).

d. Topic Representation

Word frequencies within each cluster were tabulated by concatenating its documents into one combined document (Grootendorst, 2022), forming the class-based term frequency-inverse document frequency (c-TF-IDF) representation that identifies each cluster's most representative terms:

$$w_{x,c} = tf_{x,c} \cdot \log\left(\frac{A}{f_x}\right) \quad (7)$$

where $w_{x,c}$ is the c-TF-IDF weight of term x in topic c , $tf_{x,c}$ its frequency within topic c , f_x its frequency across the whole corpus, and A is the average document length across all clusters. Higher $w_{x,c}$ marks terms that are frequent within one cluster but rare in others.

5. Hyperparameter Grid Search

Prior to optimization, a default BERTopic configuration was executed for each preprocessing scenario to establish a within-framework baseline, allowing the effects of hyperparameter optimization to be distinguished from those of preprocessing. Grid search was employed because it systematically evaluates all parameter combinations within a predefined search space, ensuring reproducibility and transparency for the four HDBSCAN hyperparameters considered (Bergstra & Bengio, 2012). The search included *min_cluster_size*, *min_samples*, *cluster_selection_method*, and *cluster_selection_epsilon*, as summarized in Table 2.

Table 2. HDBSCAN Hyperparameter Search Space

Parameter	Description	Values Tested
<i>min_cluster_size</i>	Minimum documents required to form a cluster	300, 400, 500, 600, 700, 800, 1000
<i>min_samples</i>	Density threshold for core-point classification	1, 3, 5
<i>cluster_selection_method</i>	Strategy for extracting flat clusters from the hierarchy	<i>eom</i> , <i>leaf</i>
<i>cluster_selection_epsilon</i>	Distance threshold for merging adjacent clusters	0, 0.02, 0.05

The search ranges were determined based on corpus characteristics and HDBSCAN recommendations rather than selected arbitrarily. Following Grootendorst (2022), *min_cluster_size* was varied from 300 to 1000 (approximately 1.4%–4.7% of the 21,298-document corpus) to obtain interpretable topics while avoiding excessive fragmentation. *min_samples* was limited to 1, 3, and 5 to reduce excessive outlier assignments, whereas both *eom* and *leaf* extraction strategies were evaluated. Finally, *cluster_selection_epsilon* was restricted to 0, 0.02, and 0.05, as larger values tend to merge distinct clusters and reduce topic separation.

6. Evaluation Framework

Model performance was assessed using four quantitative metrics, complemented by qualitative visual inspection, capturing both the statistical and practical dimensions of topic quality. Topic coherence was measured using the C_v score, the average pairwise semantic similarity between top words of each topic, itself averaged across all topics (Röder et al., 2015):

$$C_v = \frac{1}{|T|} \sum_{t \in T} \frac{1}{|S_t|} \sum_{(w_i, w_j) \in S_t} \text{sim}(w_i, w_j) \quad (8)$$

with T the set of generated topics, S_t the word pairs within topic t , and $\text{sim}(w_i, w_j)$ the cosine similarity between word context vectors derived from normalized pointwise mutual information (NPMI). Higher C_v indicates stronger semantic association within a topic.

Topic uniqueness weights each keyword by the inverse of its frequency across all topics (Nan et al., 2019):

$$TU = \frac{1}{K} \sum_{k=1}^K \frac{1}{|T_k|} \sum_w \frac{1}{df(w)} \quad (9)$$

with K the total number of topics, T_k the top keywords of topic k , and $df(w)$ the number of topics containing keyword w . Higher values indicate a more thematically distinct topic set, since shared keywords are down-weighted.

Document coverage measures the proportion of reference documents containing at least one topic keyword (Doi et al., 2024):

$$DC = \frac{|\{d_{ref}: d_{ref} \text{ contains } \geq 1 w_{topic}\}|}{|d_{ref}|} \quad (10)$$

with d_{ref} a document in the reference corpus and w_{topic} the set of generated topic keywords. Higher DC indicates topics that are not only coherent but representative of the corpus as a whole.

Factuality assesses the proportion of topic keywords actually grounded in the source corpus, since embedding-based models can generate keywords not literally present in the documents they summarize (Doi et al., 2024):

$$Fa = \frac{|\{w_{topic}: w_{topic} \text{ present in } \geq 1 d_{ref}\}|}{|w_{topic}|} \quad (11)$$

with w_{topic} and d_{ref} as defined above. Values close to 1 indicate that nearly all topic keywords are genuinely drawn from the corpus.

For each scenario, the four metrics computed across all 126 configurations (outlier ratio sign-reversed so that higher is always better) were standardized and analyzed with Principal Component Analysis to explore correlations and trade-offs among them. Candidates with competitive quantitative scores, a sufficient and interpretable number of topics, and a low outlier ratio were then examined qualitatively via the intertopic distance map (Schneijderberg et al., 2026), and the final configuration for each scenario was selected on this joint basis before comparison against the default BERTopic model and the LDA baseline. For LDA, the analogous qualitative check used LDAvis.

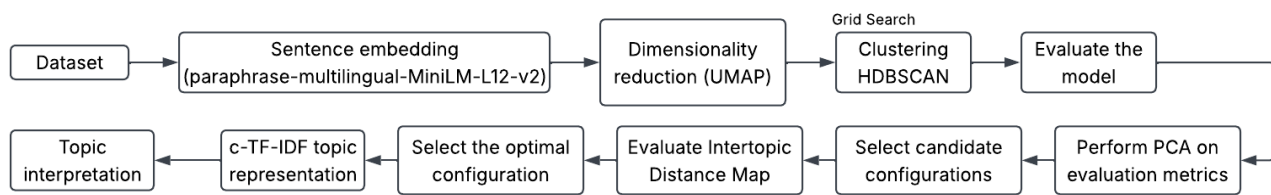


Figure 1. Research Framework for BERTopic Hyperparameter Optimization

C. RESULT AND DISCUSSION

1. Preprocessing Outcomes

Prior to topic modelling, the raw corpus of 21,298 posts was transformed through three preprocessing scenarios, each applying a different combination of token filtering and word-merging strategies as described in the Methods section. Table 3 summarizes the resulting dataset characteristics, which provide the statistical basis for interpreting the topic modelling results in the following subsections.

Table 3. Statistical Characteristics of the Preprocessed Datasets

Dataset	Documents	Avg. Doc. Length	Vocabulary Size	Unique Tokens
Raw Data	21,298	18.377	56,687	391,390
Scenario 1	21,285	8.648	56,687	184,080
Scenario 2	20,977	8.336	19,483	174,859
Scenario 3	21,283	8.002	20,481	170,303

Scenario 2 produced the smallest vocabulary and token count, a direct statistical consequence of removing numerical and single-character tokens, while Scenario 3 yielded the shortest average document length as a result of word-merging operations that collapsed compound expressions into single tokens. These distributional differences in corpus composition provide the statistical basis for the divergent topic structures discussed in the following subsections.

2. Default BERTopic Performance

BERTopic was first evaluated under its default HDBSCAN configuration for each scenario, establishing a baseline that isolates the effect of hyperparameter optimization from preprocessing. Table 4 presents the resulting evaluation.

Table 4. Evaluation Results of BERTopic with Default Configuration

Dataset	N Topics	Coherence	Uniqueness	Coverage	Factuality
Scenario 1	350	0.383	0.829	0.992	1.000
Scenario 2	334	0.380	0.827	0.994	1.000
Scenario 3	367	0.374	0.831	0.990	1.000

Default HDBSCAN settings produced between 334 and 367 topics per scenario, an order of magnitude beyond what is practically interpretable, despite coherence, coverage, and factuality all appearing statistically acceptable in isolation. Because these metrics are computed per topic or per document rather than as a function of topic count, none of them signals over-fragmentation directly, which only the raw topic count and the outlier ratio expose.

This fragmentation is illustrated concretely in Table 5, which lists representative topics generated under the default configuration for Scenario 1. Alongside a broad, high-count topic capturing general admission and campus-life discourse (Topic 0), the default configuration also produced narrowly defined topics built around only a handful of co-occurring keywords, such as Topic 349, defined almost entirely by informal expressions like “ngarahin”, “patah”, and “hahhahaha”, with only 10 assigned documents. Such topics reflect idiosyncratic local variation rather than a distinct, interpretable theme, exemplifying the over-partitioning behavior that motivates hyperparameter optimization.

Table 5. Example Topics Generated by BERTopic with Default Configuration (Scenario 1)

Topic	Count	Top Keywords	Interpretation
-1	7,235	<i>kampus, kuliah, ipb, terima, sm, masuk, ...</i>	Outlier cluster (unassigned documents)
0	556	<i>alumni, sertifi, 1000, domisili, akreditasi, ...</i>	SNBP admission strategy
1	393	<i>lulus, d3, kerja, akpol, telat, exchange, ...</i>	Mixed, no coherent single theme
2	363	<i>mahasiswa, ojol, beri, sampah, ...</i>	Mixed, no coherent single theme
...	...	...	...
349	10	<i>ngarahin, patah, ngafalin, hahhahaha, ...</i>	Mixed, no coherent single theme

3. HDBSCAN Hyperparameter Optimization

Model selection followed the two-stage statistical protocol described in the Methods section, combining standardization and Principal Component Analysis (PCA) of the four quantitative metrics with qualitative inspection of the intertopic distance map.

a. PCA Exploration

For a metric x , each configuration’s score across the 126 candidates within a scenario was first standardized as:

$$z = \frac{x - \mu}{\sigma} \quad (12)$$

with μ and σ the mean and standard deviation across configurations, placing coherence, uniqueness, coverage, and the sign-reversed outlier ratio on a common scale before PCA. The first two principal components jointly explained 95.7% of total variance in Scenario 1, 91.5% in Scenario 2, and 86.0% in Scenario 3, confirming that a two-dimensional projection captured the dominant structure of configuration quality. Figure 2 shows the resulting biplots.

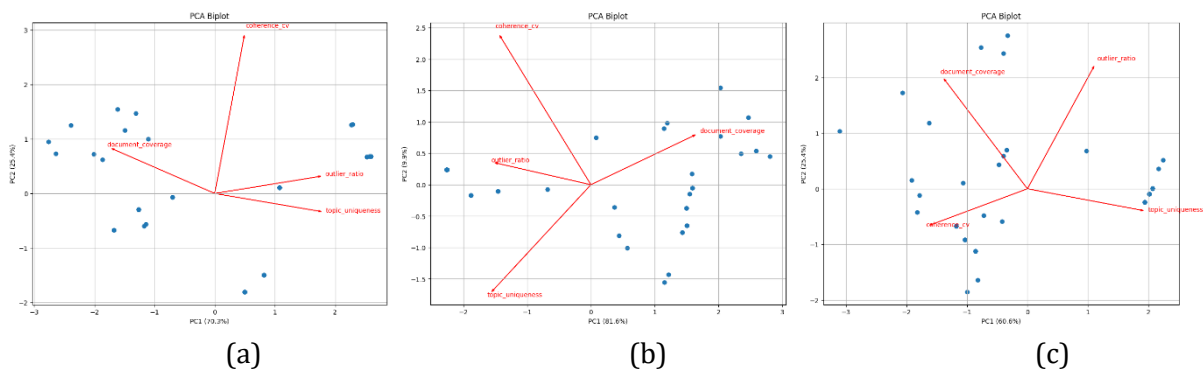


Figure 2. PCA biplots of standardized evaluation metrics across 126 HDBSCAN configurations for (a) Scenario 1; (b) Scenario 2; (c) Scenario 3.

The loading structure differed across scenarios. In Scenarios 1 and 2, *coherence_cv* formed a near-orthogonal angle with *topic_uniqueness*, indicating largely independent dimensions of configuration quality. In Scenario 3, *coherence_cv* pointed closer to directly opposing *topic_uniqueness*, indicating a sharper trade-off under the most aggressive preprocessing strategy. *Outlier_ratio* and *topic_uniqueness* pointed in broadly similar directions in Scenarios 1 and 3, suggesting that more thematically distinctive configurations also tended to leave more documents unassigned. This trade-off is why a multi-criteria selection rule was necessary.

b. Configuration Selection Process

A shortlist of the three most statistically competitive HDBSCAN configurations per scenario was drawn using the PCA-informed trade-off structure, balancing coherence, uniqueness, and coverage against a low outlier ratio. Table 6 lists these shortlisted candidates.

Table 6. Top-Performing HDBSCAN Configurations Shortlisted per Scenario

Dataset	MCS	MS	Method	Eps.	K	Coherence	Uniqueness	Coverage
Scenario 1	300	5	<i>eom</i>	0.02	3	0.634	0.867	0.896
	300	3	<i>eom</i>	0.02	3	0.634	0.867	0.896
	400	5	<i>eom</i>	0.05	2	0.615	0.900	0.894
	400	5	<i>leaf</i>	0.05	9	0.623	0.767	0.922
	500	5	<i>leaf</i>	0.02	7	0.608	0.786	0.919
Scenario 2	500	3	<i>eom</i>	0.02	2	0.636	0.950	0.899
	400	3	<i>eom</i>	0.00	3	0.600	0.900	0.903
	300	3	<i>eom</i>	0.05	6	0.590	0.783	0.920
	300	1	<i>eom</i>	0.02	6	0.523	0.850	0.922
	500	5	<i>leaf</i>	0.02	7	0.600	0.771	0.925
Scenario 3	400	1	<i>eom</i>	0.05	6	0.569	0.767	0.902
	600	3	<i>leaf</i>	0.02	5	0.538	0.760	0.904
	500	1	<i>eom</i>	0.05	5	0.522	0.800	0.900
	400	5	<i>leaf</i>	0.00	8	0.575	0.750	0.909
	1000	3	<i>leaf</i>	0.00	4	0.565	0.775	0.895

Note. MCS = *min_cluster_size*; MS = *min_samples*; Eps. = *cluster_selection_epsilon*; K = *number of topics*. *Factuality* equaled 1.0 for all listed configurations and is omitted.

The final configuration was not simply the highest-coherence candidate in every scenario. For Scenario 1, [300, 5, *eom*, 0.02] was selected over a tied candidate for its stricter density requirement. For Scenario 2, the highest-coherence candidate produced only two topics, so [300, 3, *eom*, 0.05] was selected instead, trading a 7.2% reduction in coherence for six interpretable topics. For Scenario 3, [500, 1, *eom*, 0.05] was selected over the highest-coherence candidate for its more favorable topic uniqueness (0.800 versus 0.767). Figure 2 confirms this pattern for Scenario 1, visually consistent with Scenarios 2 and 3.

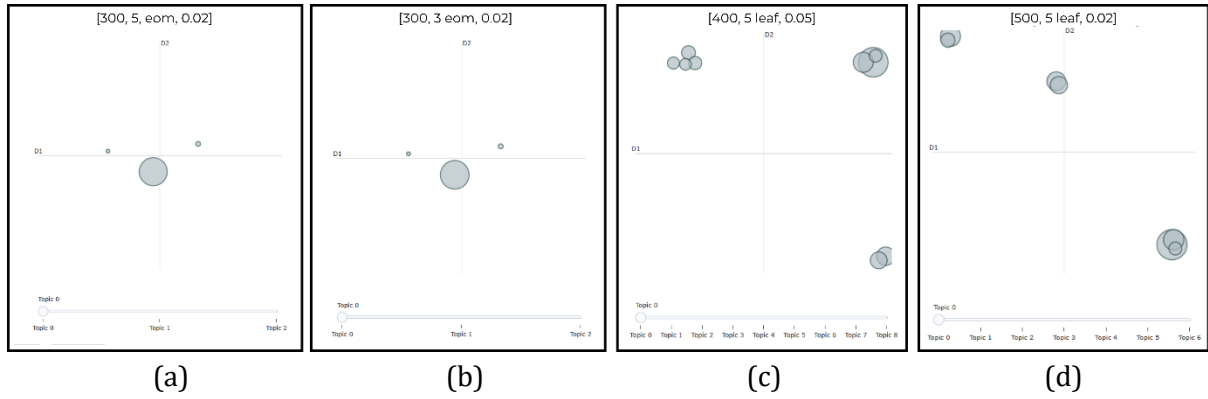


Figure 3. Intertopic distance maps of shortlisted HDBSCAN configurations for Scenario 1: (a) [300, 5, *eom*, 0.02]; (b) [300, 3, *eom*, 0.02]; (c) [400, 5, *leaf*, 0.05]; (d) [500, 5, *leaf*, 0.02].

The *eom* configurations consistently produced compact, well-separated clusters, while the *leaf* configurations produced visibly more numerous topics positioned in closer proximity, reinforcing the quantitative selection above. The proportion of documents left unassigned by HDBSCAN, central to interpreting this behavior but not applicable to LDA, is given by:

$$\text{Outlier Ratio} = \frac{|\{d: \text{label}(d) = -1\}|}{|D|} \quad (13)$$

with D the full document set. Table 7 summarizes the final configuration selected for each scenario together with its resulting outlier ratio.

Table 7. Selected Optimal HDBSCAN Configurations per Scenario

Dataset	MCS	MS	Method	Eps.	K	Coherence	Outlier %	Uniqueness
Scenario 1	300	5	<i>eom</i>	0.02	3	0.634	11.3	0.867
Scenario 2	300	3	<i>eom</i>	0.05	6	0.590	19.2	0.783
Scenario 3	500	1	<i>eom</i>	0.05	5	0.522	38.0	0.800

Note. MCS = *min_cluster_size*; MS = *min_samples*; Eps. = *cluster_selection_epsilon*; K = *number of topics*; Outlier % computed via Equation (13).

Relative to the default configuration, this optimization increased coherence by 40% to 66% while collapsing the topic count from the hundreds to between three and six interpretable themes, confirming that default HDBSCAN settings over-partition the embedding space into numerous small, statistically unstable clusters (Janssens et al., 2025). This gain is directionally consistent with Borčín & Jose (2024), who found that departing from default settings materially altered coherence, yet it exceeds the more incremental gains of Feng et al. (2025), plausibly because this corpus required a far larger optimal minimum cluster size and because preprocessing was tuned jointly with clustering rather than held fixed.

4. Statistical Interaction Between Preprocessing and HDBSCAN Hyperparameters

Table 6 makes it possible to isolate the statistical contribution of each HDBSCAN hyperparameter individually. *cluster_selection_method* was the most influential parameter, as across all three scenarios, *eom* candidates consistently produced outlier ratios between 11% and 38%, whereas *leaf* candidates exceeded 50% in most configurations, for example rising from an outlier ratio comparable to *eom* at low topic counts to substantially higher values as *leaf* pushed topic counts to 7–9 in Scenarios 1 and 2. This confirms *eom*'s tendency to favor larger and statistically more stable clusters by maximizing total cluster stability, Equation (6), across the condensed tree, rather than extracting the finest-grained partitions the tree can support (McInnes et al., 2020; Janssens et al., 2025).

min_cluster_size was the second most influential parameter, though its optimal value differed by scenario, remaining at 300 for Scenarios 1–2 but rising to 500 for Scenario 3, indicating that the statistically optimal density threshold depends on the embedding structure each preprocessing strategy produces rather than on a single fixed default (Borčin & Jose, 2024). By contrast, *min_samples* showed only a marginal effect in isolation, as in Scenario 1, the two top-ranked configurations, [300, 5, *eom*, 0.02] and [300, 3, *eom*, 0.02], differed only in *min_samples* yet produced numerically identical coherence, uniqueness, and coverage scores, indicating that this parameter mattered only as a secondary tie-breaker once *min_cluster_size* and *cluster_selection_method* were fixed. *cluster_selection_epsilon* showed the weakest effect of all four parameters, with several shortlisted candidates in Table 6 sharing identical topic counts and coherence scores despite different epsilon values, indicating that distance-based cluster merging was statistically inconsequential for this corpus.

These parameter effects cannot be interpreted in isolation from preprocessing, because BERTopic's clustering stage operates entirely on the embedding geometry that the preceding pipeline stages produce. HDBSCAN's core distance and mutual reachability distance (Equation 5) are computed directly from the local density of points in the UMAP-reduced embedding, so any preprocessing choice that alters vocabulary and document representation propagates statistically into the very density estimates on which cluster formation, and therefore *min_cluster_size*'s effective threshold, depends (Zimmermann et al., 2024). This mechanism explains why *min_cluster_size* shifted from 300 to 500 between Scenarios 1–2 and Scenario 3, and why coherence and the outlier ratio moved together with preprocessing intensity across Table 7, from 0.634 and 11.3% in Scenario 1 to 0.590 and 19.2% in Scenario 2 and 0.522 and 38.0% in Scenario 3.

Word-merging in Scenario 3 collapses compound expressions into single tokens, thinning the local neighborhood density that UMAP and HDBSCAN rely on for stable cluster formation, so that more aggressive preprocessing improved topic uniqueness only at a direct statistical cost to coherence and document coverage. This cost is also visible in the PCA loading structure in Figure 2, where coherence_cv and topic_uniqueness shifted from a near-orthogonal to a near-opposing relationship between Scenarios 1 and 3, indicating that the trade-off between these two metrics intensified precisely as preprocessing became more aggressive. Taken together, these patterns confirm that preprocessing decisions and HDBSCAN hyperparameters function as statistically interdependent determinants of BERTopic's topic quality (Farea et al., 2024). Practically, a preprocessing pipeline cannot be finalized before clustering is tuned, since a

configuration transferred unchanged from a lighter to a more aggressive pipeline risks a sharp rise in the outlier ratio, as observed here from 11.3% in Scenario 1 to 38.0% in Scenario 3.

5. LDA Baseline Performance

LDA was evaluated under the same three preprocessing scenarios. The optimal number of topics K was selected as the value maximizing coherence (Equation 8), as shown in Figure 4.

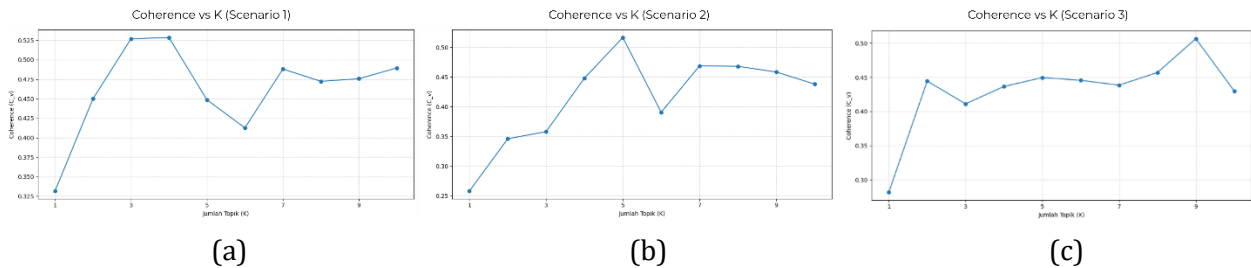


Figure 4. Topic coherence as a function of the number of topics K for LDA under (a) Scenario 1; (b) Scenario 2; (c) Scenario 3.

Figure 4 shows that coherence rose sharply from $K = 1$ before fluctuating across a broad plateau in each scenario, with the optimal K increasing from 4 in Scenario 1 to 5 in Scenario 2 and 9 in Scenario 3. This upward shift indicates that the vocabulary changes introduced by more aggressive preprocessing required a finer-grained topic structure for LDA to reach its best attainable coherence, even though that coherence still declined slightly across scenarios, from 0.528 to 0.506, the reverse of the pattern observed for optimized BERTopic in Table 7. Table 8 summarizes the resulting evaluation at each scenario's optimal K .

Table 8. LDA Baseline Evaluation Results per Scenario

Dataset	K	Coherence	Uniqueness	Coverage	Factuality
Scenario 1	4	0.528	0.945	0.968	1.000
Scenario 2	5	0.517	0.980	0.960	1.000
Scenario 3	9	0.506	0.982	0.980	1.000

6. Comparison Between Optimized BERTopic and LDA

Optimized BERTopic achieved coherence scores between 0.522 and 0.634 across the three scenarios. This consistently exceeded LDA's range of 0.506 to 0.528. This advantage is statistically attributable to BERTopic's use of contextual embeddings rather than word co-occurrence statistics for measuring document similarity (Egger & Yu, 2022; Babalola et al., 2024). Conversely, LDA achieved higher topic uniqueness (0.945–0.982 versus 0.783–0.867) and document coverage (0.960–0.980 versus 0.896–0.921). This reflects a structural difference between the two models rather than a quality difference alone. LDA assigns every document a topic-proportion vector across all K topics. BERTopic's HDBSCAN component, by contrast, explicitly labels low-density documents as statistical outliers (Equation 13), trading document coverage for tighter, more separable topic clusters. Neither pattern indicates a shortcoming of either model, but rather a genuine statistical trade-off between semantic precision and corpus-wide representativeness. For institutional reputation monitoring this trade-off has a concrete

reading, as BERTopic's higher coherence suits surfacing a small set of clearly separated, actionable concerns, whereas LDA's near-complete coverage suits an exhaustive audit in which no complaint should be dropped, as shown in Figure 5.

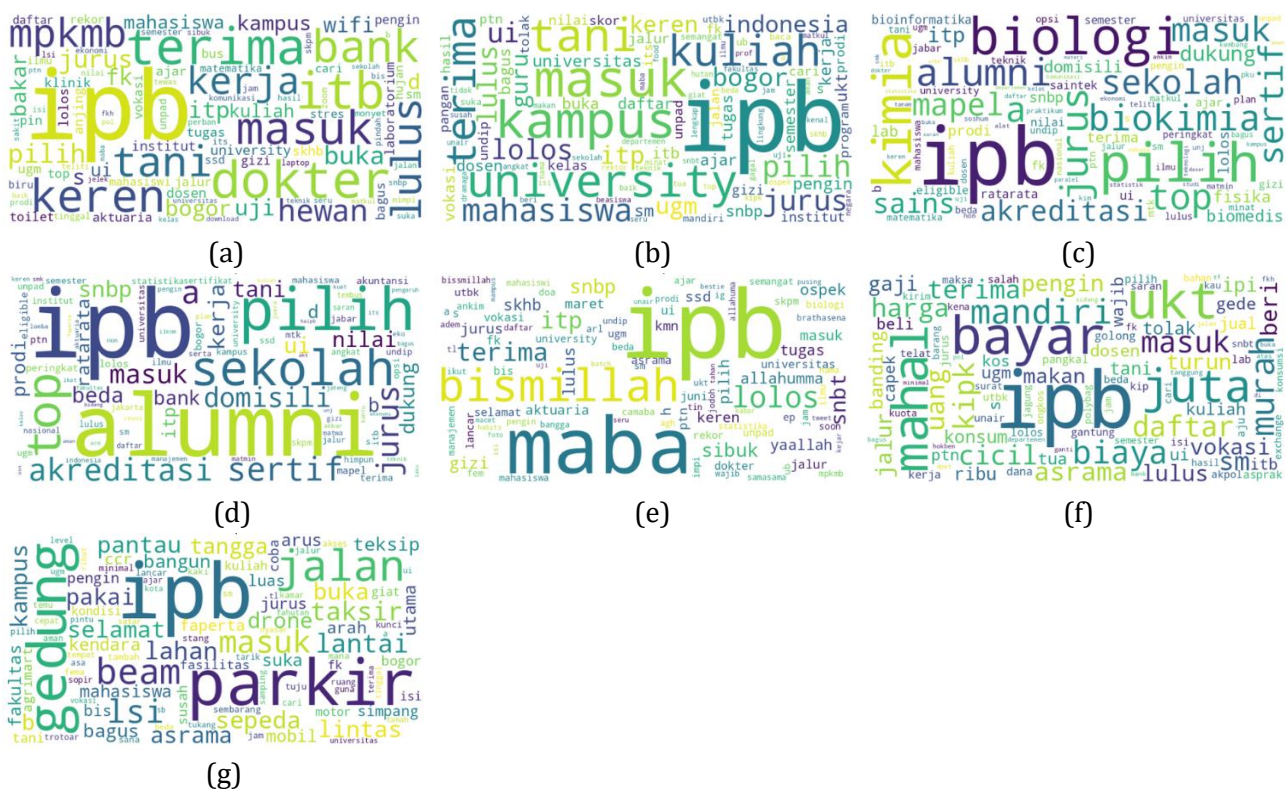


Figure 5. Word clouds of the seven topics identified by the optimized BERTopic model for Scenario 2 weighted by c-TF-IDF (Equation 7). (a) Topic -1 (Outlier); (b) Topic 0; (c) Topic 1; (d) Topic 2; (e) Topic 3; (f) Topic 4; (g) Topic 5.

Scenario 2 is used to illustrate this comparison qualitatively, as it occupies the statistical middle ground exposed in Section 4. Scenario 1's lighter preprocessing yielded only three broad topics, while Scenario 3's aggressive word-merging sharpened uniqueness but left 38.0% of documents unassigned (Table 7), so Scenario 2's six topics and moderate 19.2% outlier ratio offer the clearest basis for comparing how the two models structure the same corpus.

Figure 5 shows that BERTopic's six topics are each anchored by a distinct, largely non-overlapping vocabulary. Admission and campus-life terms ("ipb", "terima", "kuliah", "universitas") are concentrated in Topic 0, while tuition and financial terms ("ukt", "bayar", "mahal", "juta") are concentrated separately in Topic 4. Topics 1–3 each capture a single further theme, academic disciplines ("biologi", "kimia"), alumni certification ("alumni", "sertif"), and admission-cycle sentiment ("maba", "bismillah"), without visibly bleeding into one another. This clean separation is consistent with c-TF-IDF weighting (Equation 7), which derives representative terms from documents already grouped by semantic similarity, so that the resulting vocabulary reinforces rather than blurs each topic's boundary, as shown in Figure 6.

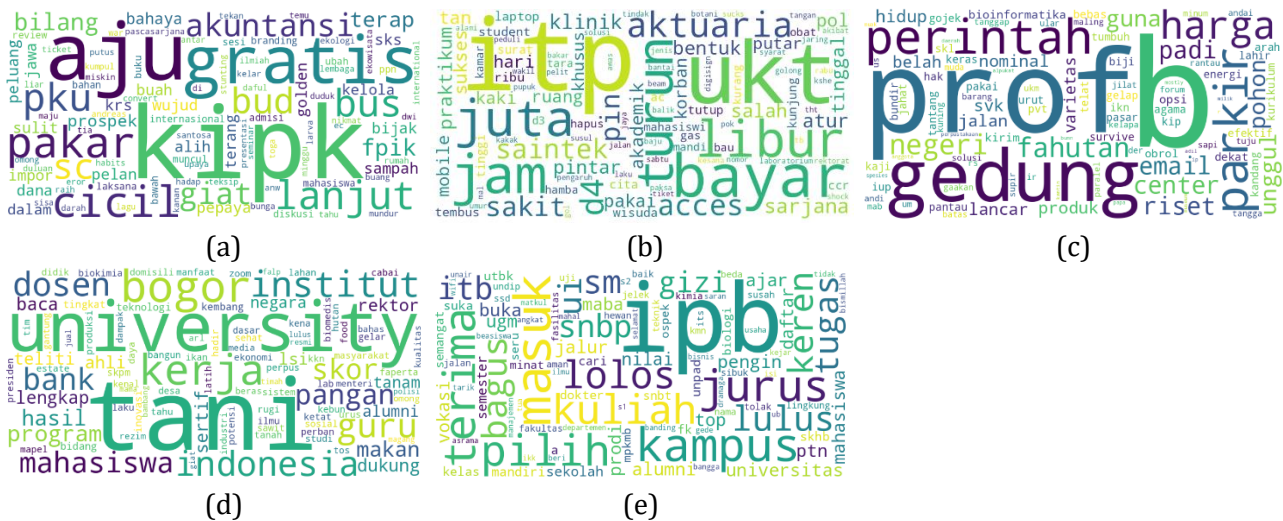


Figure 6. Word clouds of the five topics identified by the LDA baseline for Scenario 2, weighted by word-topic probability. (a) Topic 1; (b) Topic 2; (c) Topic 3; (d) Topic 4; (e) Topic 5.

Figure 6 shows a less clean separation for LDA. The cost-related discourse that BERTopic consolidated into a single topic is instead split across two LDA topics with related content, Topic 0 ("kip", "gratis", "cicil", "akuntansi," concerning KIP-K scholarships) and Topic 1 ("ukt", "bayar", "juta", "saintek", concerning tuition payment), while Topic 2 mixes campus facilities, parking, and pricing terms into a single, unfocused topic. This apparent contradiction, LDA scoring higher on topic uniqueness (Equation 9) despite these related topics, is statistically informative rather than an inconsistency, since uniqueness is computed from literal keyword overlap rather than semantic content. LDA's Topic 0 and Topic 1 draw on largely disjoint vocabulary, so despite both concerning the cost of studying at IPB, they register as lexically unique while remaining thematically related.

BERTopic's embedding-based clustering, by contrast, groups documents by semantic similarity before extracting representative terms, which is precisely why its topics achieve higher coherence but can share more overlapping vocabulary. This distinction, consistent with evidence that sparse word co-occurrence limits LDA's capacity to separate semantically related topics in short text (Albalawi et al., 2020; Zimmermann et al., 2024), indicates that BERTopic's semantic clustering produced more thematically coherent topics than LDA's distributional approach, in line with prior comparative evidence on short, informal text (Egger & Yu, 2022; Mutsaddi et al., 2025; Medvecki et al., 2024).

7. Dominant Themes and Implications for Indonesian Social Media Discourse

Comparing the word clouds of both models for Scenario 2 (Figures 6–7) reveals three themes that recur consistently regardless of clustering approach, admission and campus life, tuition and financial concerns, and campus facilities. Their independent emergence in both BERTopic's embedding-based and LDA's co-occurrence-based approach lends cross-model support to their prominence in this discourse. Each model also surfaced themes the other did not resolve as distinctly. BERTopic further separated alumni certification and admission-cycle sentiment into their own topics, while LDA's Topic 3 foregrounded IPB University's identity as an agricultural institution, consistent with its founding mandate in food security and agricultural sciences. Together, these results indicate that public discourse about IPB

University on X is anchored around practical, admission- and cost-related concerns, with institutional identity forming a secondary but recurring layer.

Beyond this thematic content, the three scenarios traced a clear progression, from Scenario 1's coarse-grained reading to Scenario 2's six diverse topics at a moderate outlier ratio and Scenario 3's sharper topics at the cost of leaving 38.0% of documents unassigned. These findings extend prior evidence of BERTopic's effectiveness on non-English short text (Mutsaddi et al., 2025; Medvecki et al., 2024) to Indonesian institutional discourse, and indicate that the joint optimization proposed in this study offers a replicable framework for extracting interpretable themes from short, informal text, while confirming that method choice should be guided by whether thematic precision or complete coverage is the primary objective (Alangari & Algethami, 2024).

8. Limitations and Threats to Validity

Several boundaries constrain how far these findings generalize. The analysis rested on a single corpus from one institution and one platform, so the specific optimal values reported here, such as the shift of *min_cluster_size* from 300 to 500, may not transfer to corpora of different size or register. The keyword-based collection may itself bias the corpus toward admission- and cost-related discourse, which should be considered before reading those themes as a complete picture of institutional reputation. Model selection relied on grid search over a bounded four-parameter space, which guarantees reproducibility but cannot rule out stronger configurations outside the tested ranges. The comparisons also rested on point estimates without significance testing, so closely ranked configurations such as the tied candidates in Table 6 should be read as practically rather than statistically distinguishable. Finally, a single multilingual embedding model was used, and the outlier-labeling mechanism makes the method least suitable where every document must be assigned a topic.

D. CONCLUSION AND SUGGESTIONS

This study developed a statistically grounded framework for jointly optimizing text preprocessing and HDBSCAN hyperparameters in BERTopic, applied to public discourse about IPB University on X. Preprocessing and HDBSCAN hyperparameters were found to interact statistically rather than independently. Word-merging in Scenario 3 thinned the local density of the UMAP-reduced embedding, shifting the optimal *min_cluster_size* from 300 to 500, raising the outlier ratio from 11.3% to 38.0%, and lowering coherence from 0.634 to 0.522. Among the four hyperparameters tested, *cluster_selection_method* had the strongest effect, with *eom* consistently outperforming *leaf*. Benchmarked against a default configuration, the optimized BERTopic model raised coherence by 40% to 66% while reducing the topic count from the hundreds to three to six interpretable themes. Compared against LDA, optimized BERTopic achieved higher coherence (0.522–0.634 versus 0.506–0.528), while LDA achieved higher topic uniqueness and coverage, a trade-off that is structural rather than a difference in quality, since LDA assigns every document to all K topics while BERTopic withholds low-density documents as outliers to produce more thematically coherent, non-overlapping topics.

Across both models, the dominant themes were admission and campus life, tuition and financial concerns, and campus facilities, with BERTopic further isolating alumni certification

and admission-cycle sentiment, and LDA foregrounding IPB University's identity as an agricultural institution. Public discourse about IPB University on X is therefore anchored primarily around practical, admission and cost-related concerns, with institutional identity forming a secondary but recurring layer.

The primary contribution of this study is methodological. It shows that in density-based neural topic modeling, preprocessing and clustering hyperparameters form a single coupled search space, because the optimal clustering configuration depends on the embedding geometry that preprocessing produces. This coupling is a property of the BERTopic pipeline rather than of this dataset, which implies that studies reporting results under one fixed preprocessing pipeline may understate how sensitive their topic structures are to configuration. Practically, the framework gives analysts and higher education institutions a reproducible route from short, informal Indonesian text to a small set of interpretable themes that can guide data-driven reputation management. Future research could extend the hyperparameter search using Bayesian or other adaptive strategies to cover a wider parameter space at lower computational cost (Bergstra & Bengio, 2012). Testing the framework on other institutions, discourse domains, and languages, and against alternative density-based or hierarchical clustering algorithms, would establish how far the observed coupling between preprocessing and clustering generalizes. Incorporating large language models for automated topic labeling would further improve interpretability and practical value.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Department of Statistics and Data Science, IPB University, for the academic guidance and research facilities extended throughout the course of this study. During manuscript preparation, generative AI tools, specifically Claude (Anthropic), ChatGPT (OpenAI), and Grammarly, were used to support language refinement, formatting consistency, and readability improvements. Their role was strictly limited to editorial assistance and did not extend to data analysis, statistical modeling, interpretation of findings, or the formulation of conclusions, all of which remain the sole intellectual work of the authors. The authors affirm full accountability for the scientific validity and originality of this manuscript.

REFERENCES

- Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. In *Information Systems*, 112, Article 102131. <https://doi.org/10.1016/j.is.2022.102131>
- Alangari, H., & Algethami, N. (2024). Exploring the effects of pre-processing techniques on topic modeling of an Arabic news article data set. *Applied Sciences*, 14(23), Article 11350. <https://doi.org/10.3390/app142311350>
- Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3, Article 42. <https://doi.org/10.3389/frai.2020.00042>
- Babalola, O., Ojokoh, B., & Boyinbode, O. (2024). Comprehensive evaluation of LDA, NMF, and BERTopic's performance on news headline topic modeling. *Journal of Computing Theories and Applications*, 2(2), 268–289. <https://doi.org/10.62411/jcta.11635>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281–305.

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Borčin, M., & Jose, J. M. (2024). Optimizing BERTopic: Analysis and reproducibility study of parameter influences on topic modeling. In *Lecture Notes in Computer Science* (Vol. 14611, pp. 147–160). Springer. https://doi.org/10.1007/978-3-031-56066-8_14
- Brandon, D. (2023). Data mining Twitter for COVID-19 sentiments concerning college online education. *Future Business Journal*, 9, Article 104. <https://doi.org/10.1186/s43093-023-00284-3>
- Capriotti, P., & Zeler, I. (2023). Analysing effective social media communication in higher education institutions. *Humanities and Social Sciences Communications*, 10, Article 656. <https://doi.org/10.1057/s41599-023-02187-8>
- Chauhan, U., & Shah, A. (2022). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys*, 54(7), Article 145. <https://doi.org/10.1145/3462478>
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (Vol. 1, pp. 4171–4186). ACL. <https://doi.org/10.18653/v1/N19-1423>
- Doi, T., Isonuma, M., & Yanaka, H. (2024). *Comprehensive evaluation of large language models for topic modeling* [Preprint]. arXiv:2406.00697. <http://arxiv.org/abs/2406.00697>
- Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7, Article 886498. <https://doi.org/10.3389/fsoc.2022.886498>
- Farea, A., Tripathi, S., Glazko, G., & Emmert-Streib, F. (2024). Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research. *Engineering Applications of Artificial Intelligence*, 136, Article 108877. <https://doi.org/10.1016/j.engappai.2024.108877>
- Feng, Y., Chen, Z., Zhang, Y., Huang, W., Zhang, X., & He, S. (2025). BERTopic_Teen: A multi-module optimization approach for short text topic modeling in adolescent health. *Frontiers in Public Health*, 13, Article 1608241. <https://doi.org/10.3389/fpubh.2025.1608241>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* [Preprint]. arXiv:2203.05794. <https://arxiv.org/abs/2203.05794>
- Janssens, W., Bogaert, M., & Van Den Poel, D. (2025). A comparative analysis of topic reduction techniques for BERTopic. *IEEE Access*, 13, 204087–204103. <https://doi.org/10.1109/ACCESS.2025.3638956>
- Laureate, C. D. P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12), 14223–14255. <https://doi.org/10.1007/s10462-023-10471-x>
- Ma, L., Chen, R., Ge, W., Rogers, P., Lyn-Cook, B., Hong, H., Tong, W., Wu, N., & Zou, W. (2025). AI-powered topic modeling: Comparing LDA and BERTopic in analyzing opioid-related cardiovascular risks in women. *Experimental Biology and Medicine*, 250, Article 10389. <https://doi.org/10.3389/ebm.2025.10389>
- McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11), Article 205. <https://doi.org/10.21105/joss.00205>
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform manifold approximation and projection for dimension reduction* [Preprint]. arXiv:1802.03426. <http://arxiv.org/abs/1802.03426>
- Medveckí, D., Bašaragin, B., Ljajić, A., & Milošević, N. (2024). Multilingual transformer and BERTopic for short text topic modeling: The case of Serbian. In *Lecture Notes in Networks and Systems* (Vol. 872, pp. 161–173). Springer. https://doi.org/10.1007/978-3-031-50755-7_16
- Mutsaddi, A., Jamkhande, A., Thakre, A., & Haribhakta, Y. (2025). BERTopic for topic modeling of Hindi short texts: A comparative study. In *Proceedings of the First Workshop on NLP for Indo-Aryan and Dravidian Languages* (pp. 22–32). ACL.
- Nan, F., Ding, R., Nallapati, R., & Xiang, B. (2019). Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the ACL* (pp. 6345–6381). ACL. <https://doi.org/10.18653/v1/P19-1640>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 3982–3992). ACL. <https://doi.org/10.18653/v1/D19-1410>

- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). ACM. <https://doi.org/10.1145/2684822.2685324>
- Romero, J. D., Feijoo-Garcia, M. A., Nanda, G., Newell, B., & Magana, A. J. (2024). Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis. *Big Data and Cognitive Computing*, 8(10). <https://doi.org/10.3390/bdcc8100132>
- Schneijderberg, C., Wieczorek, O., & Steinhardt, I. (2026). Topic modeling with BERTopic. In *The Handbook of Qualitative and Quantitative Content Analysis* (pp. 558–592). Routledge. <https://doi.org/10.4324/9781003496397-21>
- Suhaeni, C., Mualifah, L. N. A., & Wijayanto, H. (2025a). LDA topic modeling analysis of public discourse on Indonesia's free nutritious meals program (MBG). *IJID (International Journal on Informatics for Development)*, 14(1), 587–600. <https://doi.org/10.14421/ijid.2025.5211>
- Suhaeni, C., Kamila, S. A., Fahira, F., Yusran, M., & Alfa Dito, G. (2025b). Exploring a large language model on the ChatGPT platform for Indonesian text preprocessing tasks. *Indonesian Journal of Statistics and Its Applications*, 9(1), 100–116. <https://doi.org/10.29244/ijsa.v9i1p100-116>
- Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, 94, Article 101582. <https://doi.org/10.1016/j.is.2020.101582>
- Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T., & Säfken, B. (2023). Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using Twitter data. *Computational Statistics*, 38(2), 647–674. <https://doi.org/10.1007/s00180-022-01246-z>
- Zimmermann, J., Champagne, L. E., Dickens, J. M., & Hazen, B. T. (2024). Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling. *Decision Support Systems*, 185, Article 114310. <https://doi.org/10.1016/j.dss.2024.114310>