ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

# BERTopic Analysis of Indonesian Biodiversity Policy on Social Media

Nuraisa Novia Hidayati[1] and Siti Shaleha[2]

## ABSTRACT

Indonesia, known for its diverse biodiversity, faces critical challenges such as habitat degradation and species loss. This study delves into public opinion regarding Indonesian government biodiversity policies by analyzing text data from X social media platforms. Leveraging BERTopic, an advanced topic modeling technique, we uncover nuanced topics related to biodiversity within tweets. Our research uniquely contributes by exploring diverse combinations of BERTopic parameters on Indonesian text, assessing their efficacy through coherence values and manual content evaluation. Notably, our findings highlight the optimal combination of sentence embedding, cluster model, and dimension reduction parameters, with Model 5 demonstrating the highest coherence score of 0.7733. Moreover, we elucidate the impact of outlier reduction techniques when applying BERTopic in an Indonesian context. Our study serves as a foundational model for categorizing Indonesian-language topics using BERTopic, showcasing the significance of tailored text processing techniques. We also reveal that while standard preprocessing methods enhance clustering outcomes, certain dataset characteristics, such as the inclusion of hashtags and mentions, can influence coherence differently across models. This work not only provides insights into public perceptions of biodiversity policies but also offers methodological guidance for text analysis in similar contexts.

*Article information:*

**Keywords:** Biodiversity, Public Opinion, Social Media, BERTopic, Topic Modeling

**Article history:**
Received: December 12, 2023
Revised: March 21, 2024
Accepted: May 2, 2024
Published: May 18, 2024 (Online)

**DOI:** 10.37936/ecti-cit.2024183.255058

## 1. INTRODUCTION

Indonesia, distinguished as the most biodiverse nation within the ASEAN region, is confronting substantial threats to its biodiversity. Principal factors contributing to this loss include deforestation, habitat destruction, the introduction of invasive species, and the degradation of carbon storage ecosystems. The country is also grappling with several systemic challenges in biodiversity conservation, such as diminished governmental funding, protracted processes for research permit acquisition, restricted access to global scientific literature, and elevated expenses associated with modern biodiversity inventory techniques [1].

In this study, we aim to analyze public opinions regarding Indonesian government policies on biodiversity management and the issues that arise on social media platform X by extracting topics within text data. The methodology employed is topic modeling, a machine-learning technique renowned for its efficiency in organizing and summarising extensive textual data. This approach allows for the extraction and categorization of semantic information from X conversations into distinct thematic groups. The selection of a topic modeling technique is critical when evaluating social media text, particularly Bahasa Indonesia content, which is distinguished by the substantial usage of slang and non-standard grammatical structures. Traditional approaches, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), have drawbacks in this area. LDA, despite its popularity, frequently fails to capture the intricate co-occurrence relationships and dynamic character of social media discourse, resulting in insufficient information extraction and low adaptability to text evolution over time [2][3]. NMF, on the other hand, is appropriate for shorter texts but requires preprocessed data and struggles with sparse datasets, limiting its effectiveness [3][4].

[1,2] The author is the National Research and Innovation Agency, Indonesia, E-mail: nunohida@gmail.com and siti075@brin.go.id
[1] Corresponding author: nunohida@gmail.com or nura017@brin.go.id

The new approach called BERTopic caught our attention. BERTopic is an advanced topic modeling technique that is based on the BERT framework, which is a machine learning algorithm that is highly efficient in organizing and summarising extensive textual data. BERTopic is designed to extract and categorize semantic information from text conversations into distinct thematic groups, making it particularly suitable for our research. Previous research has shown that BERTopic, an advanced topic modeling technique based on the BERT framework, outperforms traditional methods such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), particularly when embedding with Pre-Trained Arabic Language Models [5]. Another research study showed that BERTopic is flexible and provides meaningful and diverse topics compared to LDA and LSA [2]. BERTopic emerges as a more robust alternative, drawing on the BERT framework's capabilities to manage the quirks of social media language better. Its higher performance in topic modeling, as seen in scenarios such as Pre-Trained Arabic Language Models, illustrates its usefulness [5]. BERTopic, unlike LDA and NMF, excels at processing brief, unstructured text, making it perfect for capturing the diverse linguistic aspects of Bahasa Indonesia social media content. Its ability to produce relevant and diverse subjects outperforms even Latent Semantic Analysis (LSA) and LDA, making BERTopic the ideal choice for this domain [6].

Given these considerations, we use the BERTopic technique to deconstruct the dataset, hoping to reveal nuanced and precise viewpoints on biodiversity management policies in Indonesia as reflected in social media debates. We seek to determine the ideal BERTopic configuration, emphasizing the importance of a technique that can adeptly handle the intricacies of Bahasa Indonesia on social media, which are characterized by prevalent slang and non-standard grammatical structures. Our goal is to quickly extract the most relevant insights by adapting the method and parameter settings to the linguistic quirks and topical nuances of Bahasa Indonesia's social media text, ensuring the extraction of important information for our study.

## 2. RELATED WORKS

The effectiveness of traditional techniques such as Latent Dirichlet Allocation (LDA) in analyzing social media data has been called into question due to their inability to capture co-occurrence relationships and their struggles with sparse datasets, as previously discussed in the introduction. Recent studies have shown that BERTopic outperforms LDA, NMF, and Top2Vec in terms of coherence and interpretability when applied to X data [4]. LDA has been shown to struggle with short texts, often leading to the formation of broad and vague topics [7][8]. On the other hand, BERTopic has been found to excel in topic coherence and diversity when used for Arabic language processing [5] and in detecting Indonesian disinformation, improving mBERT model performance with an accuracy of 0.9051, precision of 0.9515, recall of 0.8233, and F1 score of 0.8828 [9]. The BERTopic-based technique greatly outperformed the LDA method in analyzing Italian Long COVID tales, successfully clustering 97.26% of documents and obtaining 91.97% total accuracy, indicating greater efficacy and precision [10].

Further, it was instrumental in revealing COVID-19's impact on Bulgaria's education system, highlighting the effectiveness of its clustering capabilities even with noisy datasets [11]. BERTopic also demonstrated superior performance in accuracy and efficiency when applied to X data [12]. De Groot et al. found BERTopic superior to LDA in topic coherence and diversity when applied to short multi-domain texts [13]. In a university-wide model, BERTopic HDBSCAN outperformed LDA with topic coherence and diversity scores of 0.091 and 0.880, respectively, compared to LDA's 0.031 and 0.718. The replacement of HDBSCAN with k-Means maintained performance without producing outliers. Hence, the exceptional performance and broad applicability of BERTopic underscore its potential for effective topic modeling across various contexts and data types [13]. Recent research by [14] shows that within the BERTopic framework, integrating Principal Component Analysis (PCA) for dimension reduction with K-means clustering for topic categorization considerably improves topic coherence. This synergistic strategy produced subjects with an exceptional coherence score of 0.8463, outperforming alternative topic extraction and coherence combinations and approaches.

## 3. METHODOLOGY

This study emphasizes the significance of different parameter combinations within the BERTopic model in relation to clustering brief Indonesian text data derived from the social media network X. The approach employed in this study is shown in Fig 1.

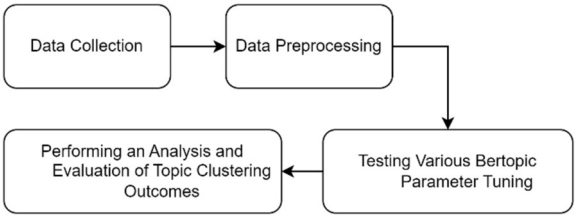


***Fig.1:*** *Research Methodology.*

### 3.1 Data Collection

The data collected originates from the social media platform X, utilizing specific keywords related

to government policies that may impact biodiversity in Indonesia. This data must encompass food biodiversity, plant biodiversity for herbal medicines, biodiversity protection, and environmental conservation. The data were gathered over three years, from 2021 to 2023, carefully selecting only relevant information and excluding advertisements or provocative statements. A total of 13,518 data points were compiled, representing public sentiment concerning biodiversity-related government policies, as explained in [15].

### 3.2 Data Preprocessing

Our data was divided into three subsets, each subjected to different preprocessing treatments referred to in our previous research. [16]. The first dataset underwent URL removal, punctuation substitution with spaces (excluding apostrophes), and replacement of non-ASCII characters with their closest ASCII equivalents. We then implemented the normalization of slang and typos, the standardization of word variations, and the reduction of the impact of slang and typographical errors. Furthermore, numerical values and non-ASCII characters were omitted, and consecutive whitespaces were reduced to a single whitespace. Using the Sastrawi library, we executed stopword removal, retaining adverbial words due to their substantial influence on sentiment; this data is called data prep 1. The second dataset underwent additional preprocessing to remove mention text, called data prep 2, and the third dataset removed both mention and hashtag text, called data prep 3. These comprehensive preprocessing procedures resulted in a fully prepared dataset comprising 13,518 data entries ready for training.

### 3.3 Testing Various BERTopic Parameter Tuning

Several factors influence BERTopic's performance in topic modeling, which might limit its usefulness. Large vocabularies are difficult to manage, as evidenced by occasions where LDA outperformed BERTopic in terms of subject quality and relevance [4][17]. The quantity of the vocabulary is an important consideration; a larger vocabulary may improve the model's accuracy but requires more computer power and resources [4]. The choice of a clustering technique is critical since it considerably influences the model's capacity to generate cohesive themes; an incorrect choice can damage the quality of the results [18]. Furthermore, the performance of BERTopic depends on the embedding model used. Different models may suit different types of data, and a poor choice might reduce topic modeling efficiency [4]. BERTopic requires particular parameter calibration to perform efficiently, with incorrect modifications potentially leading to poor results [4]. However, BERTopic's design is flexible, allowing users to customize dimensionality reduction, clustering, tokenization, and weighting schemes, as well as fine-tune representation parameters, allowing the model to be more accurately aligned with specific dataset nuances and research objectives [3]. BERTopic's topic modeling involves three crucial steps: transformer embedding, dimensionality reduction, and clustering [19]. It utilizes a transformer to generate dense vectors for textual fragments [20] . This research employs multiple scenarios, including diverse Indonesian language embeddings and customizing the clustering phase.

The BERTopic model starts by turning input documents into numbers. Due to the "all-MiniLM-L6-v2" sentence transformer's ability to capture semantic similarity well—as demonstrated by a cosine similarity of 0.7 [21]—it frequently uses it. The "distiluse-base-multilingual-cased-v2" version, which supports 50+ languages, including Indonesian, enables the identification of semantically similar sentences within or across languages [22]. The "cahya/bert-base-indonesian-522M" model, which was trained using Indonesian Wikipedia datasets and a masked language modeling (MLM) objective, is another tool for processing the Indonesian language [23]. In our research, we also employed the 'firqaaa/indo-sentence-bert-base,' an Indonesian Sentence BERT model designed for semantic similarity analysis, available in the Hugging Face repository. This model is specifically tailored for semantic analysis in the context of the Indonesian language. A critical facet of BERTopic is the dimensionality reduction of input embeddings. The study employs UMAP, a default choice in BERTopic due to its ability to capture local and global traits of high-dimensional space effectively. Despite its lack of interpretability compared to linear techniques such as PCA and NMF, UMAP outperforms visualization quality and global structure preservation with faster runtime performance. Therefore, it effectively balances global and local structure preservation, according to a comparison of dimension reduction algorithms. [24].

This research focuses on clustering techniques, particularly HDBSCAN, K-Means, and agglomerative clustering, in processing input embeddings for topic extraction. HDBSCAN is recognized for effectively capturing structures with varying densities, influenced by parameters like min_cluster_size, which was set to 15 and 80 in this study. The vectorizer (CountVectorizer) in the BERT model is crucial for topic representation generation, offering flexibility in parameter tuning. Applied prior to training, it can reduce the size of the resulting c-TF-IDF matrix [18]. BERTopic uses Bag-of-Words representation and c-TF-IDF weighting, enabling swift generation of keywords independent of the clustering process. This technique allows for easy post-training updates on topics without re-training. However, further fine-tuning of topic representations may be desir-

able. Our study used multi-aspect topic modeling to create multiple representations of a single topic using various topic representation models, specifically PartOfSpeech and KeyBERTInspired. Though not employed by default, these models offer additional fine-tuning. The following table presents the parameters that were examined in the BERTopic model.

***Table 1:*** *Model Parameter.*

| Model | Parameter |
|---|---|
| *BERTopic* 1 | embedding_model = SentenceTransformer("all-MiniLM-L6-v2") |
| | umap_model = UMAP(n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine') |
| | hdbscan_model = HDBSCAN(min_cluster_size=15, metric='euclidean', cluster_selection_method='eom', prediction_data=True) |
| | vectorizer_model = CountVectorizer() |
| | ctfidf_model = ClassTfidfTransformer() |
| | representation_model = KeyBERTInspired() |
| *BERTopic* 2 | embedding_model = SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2") |
| | dim_model = PCA(n_components = 5) |
| | cluster_model = KMeans (n_clusters =50) |
| | vectorizer_model = CountVectorizer() |
| | ctfidf_model = ClassTfidfTransformer() |
| | representation_model = KeyBERTInspired() |
| *BERTopic* 3 | embedding_model = SentenceTransformer('firqaaa/indo-sentence-bert-base') |
| | umap_model = UMAP(n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine') |
| | cluster_model = AgglomerativeClustering(n_clusters=50) |
| | vectorizer_model = CountVectorizer() |
| | ctfidf_model = ClassTfidfTransformer() |
| | representation_model = KeyBERTInspired() |
| *BERTopic* 4 | embedding_model = distiluse-base-multilingual-cased-v2 |
| | umap_model = U_MAP (n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine', random_state=42) |
| | cluster_model = HDBSCAN = min_cluster_size=15, metric='euclidean', cluster_selection_method='eom', prediction_data=True |
| *BERTopic* 5 | embedding_model = cahya/bert-base-indonesian-522M |
| | umap_model = U_MAP = n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine' |
| | cluster_model = HDBSCAN = min_cluster_size=15, metric='euclidean', cluster_selection_method='eom', prediction_data=True |
| *BERTopic* 6 | embedding_model = cahya/bert-base-indonesian-522M |
| | umap_model = U_MAP = n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine' |
| | cluster_model = HDBSCAN = min_cluster_size=15, metric='euclidean', cluster_selection_method='eom', prediction_data=True |
| | reduce$_o utliers$() |

## 3.4 Performing an Analysis of Topic Clustering Outcomes

In this study, evaluation is conducted in two ways, namely qualitative assessment and quantitative assessment.

### 3.4.1 Qualitative assessment

Qualitative assessment is done by manually examining each topic generated and checking the relevance of each keyword generated in each cluster. Then, each generated topic will be included in the data to observe its grouping in each dataset. The examination aims to ensure whether the generated topics have indeed grouped the data based on their core similarities or not.

### 3.4.2 Quantitative assessment

Quantitative assessment involves calculating the coherence value generated by each topic model. In this study, the coherence value is computed by performing ten iterations of the fit transform, meaning each model undergoes training ten times, and the coherence value is determined for each training iteration. In this study, we utilize the coherence model from Gensim to evaluate our model, particularly focusing on cross-validation (cv) coherence. This study chose to use cv coherence, which stands out as the best coherence measure because of several factors. Firstly, it combines multiple coherence metrics, including NPMI, which is known for capturing semantic relationships between words, and the boolean sliding window, which assesses word co-occurrences. This comprehensive approach likely contributes to its effectiveness in capturing the coherence of topics. Moreover, the boolean sliding window method, when utilized in cv, performs exceptionally well. This method implicitly represents distances between word tokens within large documents, allowing for a more nuanced understanding of word relationships [25] .

In the *cv* coherence approach, every word within a topic is compared to all other topics to assess its coherence. The process entails analyzing word occurrences within a sliding window consisting of 110 words, capturing both direct and indirect confirmations of word relationships. For each topic, the N most probable words are selected, and word vectors of size N are created for each word. Within these vectors, the Normalized Pointwise Mutual Information (NPMI) values between each word and every other word in the topic are stored. These word vectors are then aggregated into one vector, representing the collective word relationships within the topic. Finally, the coherence score ($C_v$ score) is calculated by averaging the cosine similarities between each topic word and its corresponding topic vector. This score provides a quantitative measure of how well the words within a topic are related to each other, thus offering insights into the coherence and interpretability of topics.

NPMI score is an advanced way to calculate the probability ($P$) of two words ($w'$ and $w''$) co-occurring in a corpus.

$$NPMI(w', w'') = \frac{\log\log \dfrac{P(w', w'') + \varepsilon}{P(w')P(w'')}}{-\log\log(P(w', w'') + \varepsilon} \quad (1)$$

$\varepsilon$ serves as a minor constant employed to prevent the calculation of a logarithm of zero. The probability is determined by analyzing a sliding window, denoted as $s$ (with $s$ equal to 110 in the case of Coherence $C_v$) and $j$ as an index of the sliding window. $D$ is document, while $d$ is document index in the corpus, which $|\delta_d|$ is the number of words in documents $d$. Then, the probabilities in the NPMI formula are calculated as follows:

$$P(w_n, w_m) = \frac{\sum_{d=1}^{D} \sum^{|\delta_d - s|} b_d, j(w_n, w_m)}{\sum_{d=1}^{D} |\delta_d| - s} \quad (2)$$

A word vector $\overrightarrow{w_{n,k}}$ length N, which is based on NPMI, is created for each topic. $w_{m,k}^{T}$ indicates a topic word at index m in topic k. This is a direct confirmation. Finally, the $C_v$ score is determined by averaging all cosine similarities. $s_{\cos}(\vec{v}, \vec{w}) = \frac{\vec{v}.\vec{w}}{|\vec{w}|x|\vec{w}|}$ Across all word indexes in the topics $(N) \times (K)$ topic index pairs.

$$c_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{\cos}(\vec{w}_{n,k}, \vec{w''}_k)}{N \times K} \quad (3)$$

## 4. RESULT AND DISCUSSION

In this study, BERTopic is used to generate topics from a collection of Indonesian-language tweets related to biodiversity policy discussion. BERTopic not only maps each word to the same topic but also produces representations of each topic. This research attempts to test BERTopic with various parameters on an Indonesian-language dataset and then assess how well the generated topics perform under different scenarios, as presented in Table 1.

The BERTopic 1 model, which is based on the all-MiniLM-L6-v2 architecture, is suited for circumstances when GPU resources are constrained, as demonstrated in three different data preparation scenarios outlined in [20]. When applied to the initial dataset ("data prep 1"), it identified 3,021 tweets as noisy or irrelevant. Among these, the top ten representative terms were "Indonesia," "Jakarta," "kebakaran" (fire), "masyarakat" (society), "ekowisata" (ecotourism), "kebijakan" (policy), "ketahanan" (resilience), "mendukung" (support), "negara" (country), and "pakai" (usage). These terms were widely spread throughout the dataset and lacked the specificity required to draw firm conclusions. The model produced 94 topics, each having a coherence score of 0.5636, indicating moderate topic consistency. However, some themes have irregularities in their top keywords. For example, the most prevalent words in subject 23 included the improper word "tai" (poop) and suspected usernames like "kkurimark" and "azwarsiregar", alongside only two relevant phrases, "waduk" (reservoir) and "wagubnya" (vice governor), implying a discussion on reservoir policy. However, the presence of irrelevant words diluted the topic's focus. This type of issue is shown in Fig 2 when there is a mismatch in topic representation, particularly in topics 23–27 and 40, where the top keywords do not adequately represent the underlying themes.

In the second data preparation scenario, referred to as "data prep 2," where mentions were omitted, we noticed an increase in the number of noisy tweets to 5,273. This was accompanied by a rise in the number of identified topics, totalling 131, and a coherence score of 0.5724. The noise-classified document featured prevalent words such as 'balita' (toddlers), 'mendukung' (support), 'Jakarta', 'makanan' (food), 'masyarakat' (society), 'kesehatan' (health), 'ekowisata' (ecotourism), 'kebakaran' (fire), and 'bahkan' (even). Anomalies were particularly noticeable in several topics; for instance, in topic 16, the word 'kendaraan' (vehicle) appeared alongside the hashtag '#iknmembawaperubahan', which pertains to the new capital city plan. This indicates a mismatch since the topics of the new capital city plan and electric vehicles are distinct. Furthermore, an abundance of unrelated hashtags was identified in topics 48, 30, and 31, making it challenging to ascertain the core theme of these topics, as shown in Fig 3.

In the third analysis scenario, termed "data prep 3," we processed tweets by eliminating mentions and hashtags, which resulted in 5,174 tweets being identified as irrelevant or 'noise'. The predominant terms in these noise tweets were 'stunting', 'Indonesia', 'balita' (meaning toddlers), 'Jakarta', 'masyarakat' (society), 'makanan' (food), 'kesehatan' (health), 'upaya' (efforts), 'keluarga' (family), and 'ketahanan' (resilience). This led to the identification of 128 distinct topics, with an average coherence score of 0.5812, indicating the consistency of the topics extracted. However, interpretative challenges arose in two specific topics: Topic 23 featured the term 'waduk' (reservoir) alongside the particular name 'mungkur' and the Javanese pronoun 'kuwi', while Topic 43 included the negation word 'tidak' (not) and the verb 'tinggalkan' (leave). These linguistic elements made it challenging to decipher the underlying themes of these two topics, as depicted in Fig 4, indicating the complexity of understanding the content within these topics.

Overall, the BERTopic 1 model demonstrates reasonable performance. Interestingly, the nature of the noise appears to be predominantly related to public health concerns, such as stunting in toddlers and food resilience, especially with the frequent mention of "makanan" (food). The study reveals that the inclusion of mentions and hashtags significantly influences the clustering process, thereby affecting the concentration on the primary subject matter.

The paraphrase-multilingual-MiniLM-L12-v2 model is used to facilitate sentence embedding in BERTopic

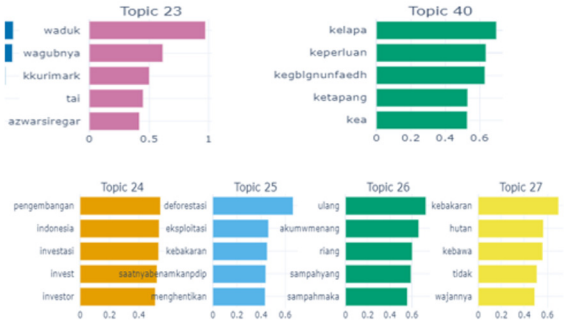


***Fig.2:*** *Anomalies Topic Generated by BERTopic 1 Model - Data Prep 1.*

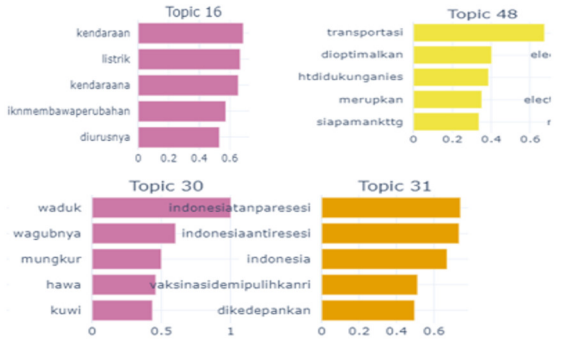


***Fig.3:*** *Anomalies Topic Generated by BERTopic 1 Model - Data Prep 2.*

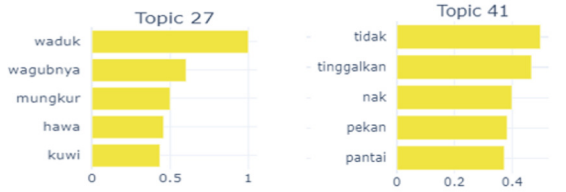


***Fig.4:*** *Anomalies Topic Generated by BERTopic 1 - Data Prep 3.*

2. This embedding technique is combined with Principal Component Analysis (PCA) to reduce dimensionality before categorizing X data into 50 discrete clusters using KMeans clustering. Comparative examination with BERTopic 1 revealed no variance in word representation among the generated subjects; however, a noticeably higher coherence value of around 0.6397 was observed. Initial data preparation (data prep 1) highlighted the dataset's tendency to cluster around hashtags related to certain events, entities, and actions in the healthcare area. Despite these repeated themes, a distinct issue with negative language arose, reflecting negative attitudes toward healthcare policies, imported staple foods, and environmental concerns. Specifically, the clustered tweet data demonstrated negative sentiments predominantly in topic 17 with the word "mahal" (expensive), topic 5 with "bodoh" (stupid) and "tolol" (idiot), topic 18 with "miskin" (poor), "mahal" (expensive), and "kurang" (lacking), and topic 13 mahal (expensive), as shown in Fig 5.

In "Data prep 2," a clear pattern emerged in the clustering of topics, particularly around policy targets, yielding a coherence score of 0.6384. This organization facilitated more coherent interpretations, as seen in topic 9, which concentrated on forest fire early warning systems through terms like "iklim" (climate), "deforestasi" (deforestation), "bencana" (disaster), "kebakaran" (fire), and "warning," indicating environmental and disaster preparedness concerns. Topic 12 focused on staple food policies, represented by words such as "beras" (rice), "pertanian" (agriculture), "petani" (farmers), and "pangan" (food), pointing to agricultural and food security discussions. Topic 15 addressed agriculture and poverty alleviation, with "pertanian" (agriculture), "subsidi" (subsidy), "kemiskinan" (poverty), and "miskin" (poor) emphasizing the economic welfare of farmers. Topic 13 highlighted herbal medicine through terms like "herbal," "obat" (medicine), and "khasiat" (efficacy), focusing on its promotion and benefits. Lastly, topic 14 delved into village governance, using words like "kampung" (village), "musyawarah" (deliberation), and "kecamatan" (sub-district), reflecting local governance and community engagement. The structured thematic grouping in "data prep 2" offered a nuanced understanding of the policy-oriented discourse within the dataset, as depicted in Fig 6.

The coherence value for the third data preparation scenario (data prep 3) for the BERTopic 2 model increased to 0.6410, indicating improved result quality and interpretability. This improvement was achieved by eliminating hashtags and mentions, allowing analysis to focus on textual content alone. The algorithm detected repeated negative sentiment issues, particularly in topic 5, which had words like 'bodoh' (stupid), 'tolol' (foolish), 'belum' (not yet), and 'bukan' (not). These terms indicate critical discussions within the dataset. Policy-related matters dominated the discussion, with a strong emphasis on rice procurement, as evidenced by phrases like 'beras' (rice), 'pertanian' (agricultural), 'petani' (farmers), 'komoditas' (commodity), and 'padi' (paddy). These reflect Indonesia's debate on agriculture policies and practices. Topic 13 focused on agricultural subsidies, denoted by the terms 'subsidi' (subsidy), 'bersubsidi' (subsidized), 'pertanian' (agriculture), 'petani' (farmers), and 'ekonomi' (economy), emphasizing economic elements and the consequences of government subsidies.

Policy-related matters dominated the discussion, with a strong emphasis on rice procurement, as evidenced by phrases like 'beras' (rice), 'pertanian' (agricultural), 'petani' (farmers), 'komoditas' (commodity), and 'padi' (paddy). These reflect Indonesia's debate on agriculture policies and practices. Topic 13 focused on agricultural subsidies, denoted by the terms 'subsidi' (subsidy), 'bersubsidi' (subsi-

dized), 'pertanian' (agriculture), 'petani' (farmers), and 'ekonomi' (economy), emphasizing economic elements and the consequences of government subsidies.

Furthermore, the model investigated additional significant themes such as coconut-derived cooking oil, represented by 'minyak' (oil),'minyakita' (oil brand), 'kelapa' (coconut), 'komoditas' (commodity), and 'goreng' (frying), representing its cultural and market relevance in Indonesia. The analysis also addressed societal and environmental issues such as stunting, climate change, herbal medicine use, electric car adoption, and forest fire challenges, which were represented by relevant keywords. Topic 48 was dedicated to Indonesia's biodiversity, with phrases such as 'Indonesia', 'tropis' (tropical), 'keanekaragaman' (diversity), and 'kelapa' (coconut), emphasizing the country's abundant biodiversity and prominence in scientific discussions.

In conclusion, the use of Principal Component Analysis (PCA) for dimension reduction and K-means clustering has been critical in discovering and categorizing keywords that reflect specific difficulties, allowing for effective differentiation across themes with similar lexical properties but different contexts. This methodological approach has been demonstrated to outperform alternative BERTopic combinations, such as PCA with HDBSCAN and UMAP with K-means, in terms of coherence scores [14]. Notably, in our research examination, BERTopic 2 witnessed a drop in coherence score, potentially due to the predefined number of topics set to 50. This fixed topic count, combined with the absence of noise clustering, led to the redistribution of tweets—previously categorized as noise in BERTopic 1—into other relevant topics, potentially affecting the overall coherence. In addition, the third data preparation scenario in this study greatly enhanced topic modeling. By removing unnecessary social media features such as hashtags and mentions, the model was able to capture the primary substance of discussions better, resulting in a more accurate and insightful analysis of various social, economic, and environmental challenges, particularly in the Indonesian context.

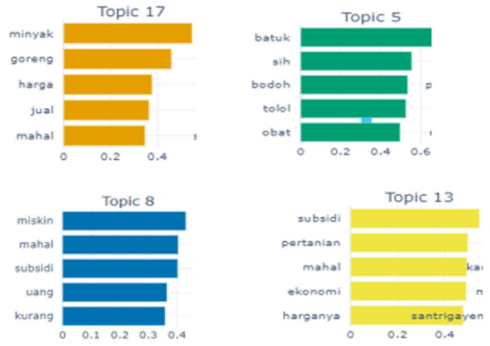


***Fig.5:*** *Negative Sentiment Topics BERTopic 2 Model - Data Prep 1.*

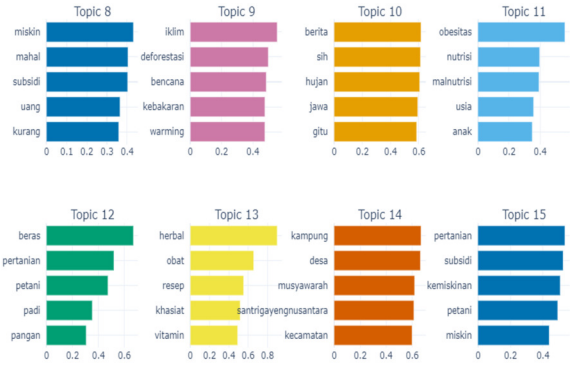


***Fig.6:*** *Topics according to policy points BERTopic 2 Model - Data Prep 2.*

BERTopic 3 for topic modeling on Indonesian tweets in this study produces more specific topic results. We used phrase embeddings trained particularly for the Indonesian language, firqaaa/indo-sentence-bert-base, in conjunction with an agglomerative clustering algorithm set to yield a maximum of 50 clusters. This method successfully addressed noisy X data, a common issue that can skew topic interpretation. It also helped to provide clear and consistent word representations across three data preparation methodologies (Data Prep 1, 2, and 3). In our data analysis, which we'll refer to as Data Prep 1, we observed convergence around several key themes, achieving a coherence score of 0.7068. Within this dataset, discussions on toddler health primarily centered on issues like "obesity" and "nutrition." Additionally, the conversation encompassed health facilities such as "PKK" and "posyandu" (community health centers). Notably, public figures like Ganjar Pranowo and President Jokowi were discussed alongside topics related to commodities such as "jagung" (corn) and "beras" (rice). Moving to Data Prep 2, which exhibited a higher coherence score of 0.7122, we noticed a stronger tendency for clustering based on hashtags. Here, the discussion prominently revolved around the promotion of electric vehicles, indicated by hashtags like "pakemobillistrikyuk" and terms like "mobil" (car) and "kendaraan" (vehicle). Moreover, political discourse emerged, featuring terms associated with figures such as the chairman of PKB, Gus Muhaimin. Finally, in Data Prep 3, which achieved a coherence score of 0.7125, distinct topics emerged that encapsulated the primary discussions within the dataset. These topics predominantly revolved around discussions concerning import regulations and farmer welfare.

The findings presented in Fig 7 indicate a high level of consistency among the results obtained using three different methods of data preparation. This consistency underscores the reliability of the BERTopic 3 model in identifying and representing the main themes within the dataset, regardless of how the data was initially processed. Notably, the topics identi-

fied cover a wide range of domains. For instance, Topic 20 encompasses terms related to agriculture, such as "sorgum" (sorghum), "jagung" (corn), and "beras" (rice). Similarly, topics related to politics include references to Ganjar Pranowo (Topic 40) and Jokowi (Topics 43 and 12). Additionally, some topics pertain to government agencies, exemplified by Topic 41, which includes terms related to the Ministry of Villages ("kemendes") and related concepts. Importantly, this thematic diversity is consistently observed across all three preprocessing approaches, indicating the robustness of the model's performance in capturing diverse themes within the dataset.

The study highlighted the effectiveness of sentence embeddings specifically trained in the Indonesian language for topic clustering. These specialized embeddings are adept at capturing the linguistic characteristics of Indonesian, thereby improving the accuracy of clustering similar topics similar to the previous study when the model employed a specific language in Arab [5]. In the comparison of clustering algorithms, HDBSCAN was noted for its ability to identify noise within the data, classifying outliers effectively. This is beneficial for the clarity of clusters but can lead to the exclusion of potentially informative data points by overly categorizing them as noise. Agglomerative clustering, in contrast, may be less noise-sensitive but compensates with its speed and consistency in generating clusters. The choice between HDBSCAN and Agglomerative clustering hinges on the specific goals of the research. If the research prioritizes operational speed and results in stability, Agglomerative clustering is advantageous [26]. However, for datasets where noise differentiation is critical, HDBSCAN's nuanced approach to noise handling may be preferred despite its potential to overlook subtle but relevant patterns within the data.

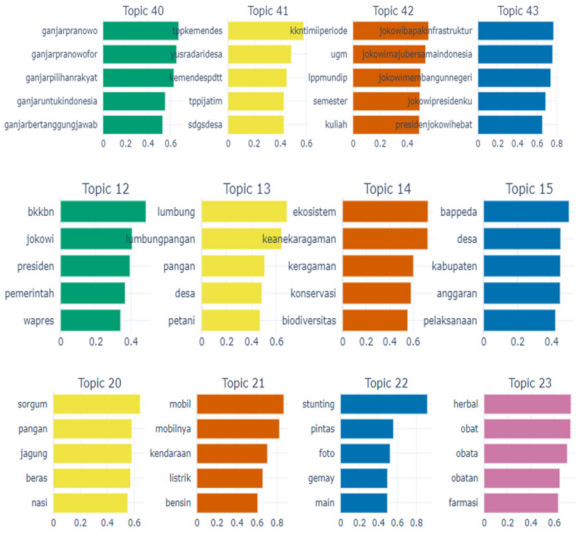


***Fig.7:*** *BERTopic 3 Topic Cluster Result.*

BERTopics 4 and 5 use different embedding models. Look at Table 1. Therefore, both produce a different number of topics, a different number of words in the outlier topic, and different topic representations. Table 2 shows the difference in the number of topics generated by both models and, of course, using different datasets. Based on Table 2, BERTopic 5 results in fewer topic clusters than BERTopic 4.

***Table 2:*** *Number of topics yang generated by BERTopic 4 dan 5, including the outlier cluster.*

| | **Prep 1** | **Prep 2** | **Prep 3** |
|---|---|---|---|
| **BERTopic 4** | 102 | 126 | 134 |
| **BERTopic 5** | 78 | 91 | 77 |

In Topic 4, documents classified as noise are grouped into Topic (-1) or the noise topic. This categorization is evident across three data preparations: Data Prep 1, 2, and 3. In Data Prep 1, 3603 documents are categorized as outliers. At the same time, in Data Prep 2, 4715 are classified as outliers, and 4200 are categorized as outliers in Data Prep 3—meanwhile, BERTopic 5 groups more documents into outliers. Specifically, 6002 documents were in Data Prep 1, 5877 documents were in Data Prep 2, and 5544 documents were in Data Prep 3. Analysis indicates that BERTopic 5 tends to group more documents into the outliers category compared to BERTopic 4. Observing the number of topics generated in Table 2, BERTopic 4 yields more topics than BERTopic 5. To validate the effectiveness of this grouping, we need to examine the correlation values generated by both models.

***Table 3:*** *Coherence values of BERTopic model 4 and 5.*

| | **Prep 1** | **Prep 2** | **Prep 3** |
|---|---|---|---|
| **BERTopic 4** | 0.7193 | 0.7128 | 0.7061 |
| **BERTopic 5** | 0.7733 | 0.7612 | 0.7176 |

Based on the average values in Table 3, BERTopic 5 produces a higher coherence score compared to BERTopic 4. This indicates that the BERTopic 5 model using Indonesian language embedding is better for topic clustering in Indonesian text with the BERT model. Furthermore, among the three datasets with different preprocessing methods, it is evident that data prep 1 outperforms the others, followed by data prep 2 and data prep 3 with the lowest performance. This suggests that data prep 1, without removing hashtags and mentions, yields a better coherence value compared to data prep 2 and 3, which remove hashtags and mentions. This implies that the coherence value consistently decreases whenever information is reduced in prep 2 by removing mentions and in prep 3 by removing mentions and hashtags.

Qualitative observation is conducted by observing both models in classifying tweets into relevant topics. Here, the observation focuses on the capabili-

ties of the models, so this time, the results from several sample tweets from the model training with data prep 1 are chosen. Table 4 categorizes tweets based on BERTopic 4 and BERTopic 5. According to Table 4, BERTopic 4 is worse in classifying tweets; this can be observed in the second tweet related to accelerating the reduction of stunting, while BERTopic 5 groups this tweet into the topic of decreased rate stunting. Still, by BERTopic 4, this tweet is grouped as outliers (-1). This is because BERTopic 5 uses an Indonesian language embedding model, so the use of native Indonesian language embeddings is proven to produce better word grouping. Abuzayed, A, et al [5] also researched topics related to grouping Arabic words using Arabic word embeddings and found that Bertopic outperforms LDA and NMF models. Therefore, in this study, we try to compare the results of Indonesian language embeddings with multilingual embeddings, and it was found that Indonesian language embeddings outperform.

***Table 4:*** *The Comparations of BERTopic 4 and 5 in Classifying Tweets.*

| **Tweet** | **BERTopic 4** | **BERTopic 5** |
|---|---|---|
| Let's keep the spirit to prevent stunting (*ayo semangat mencegah stunting*) | 2. *stunting, stunting emang, cegah stunting, rembuk* (stunting, stunting indeed, prevent stunting, discuss) | 9*_cegah stunting _stunting cegah* (prevent stunting_prevent stunting) |
| Convergence to accelerate the reduction of stunting in Jepara *(konvergensi percepatan penurunan stunting jepara)* | -1 *bakar hutan, Indonesia, beras impor* (burn forest, Indonesia_rice import) | 6. *turun, angka stunting, turun stunting, angka* (decrease, stunting rate, decrease stunting, rate) |
| Raising stunting in Pariaman decreases stunting percentage *(angkat stunting pariaman turun persen stunting)* | 38*_target percentage_target decrease_prevalensi target_percentage decrease* (target persen_target turun_target prevalensi_turun persen) | 6.*turun, angka stunting, turun stunting, angka* (decrease, stunting rate, decrease stunting, rate) |
| Posyandu, let's prevent stunting *(posyandu ayo cegah stunting)* | 2.*stunting, stunting emang, cegah stunting, rembuk* (stunting, stunting indeed, prevent stunting, discuss) | 9*_cegah stunting _stunting cegah* (prevent stunting_prevent stunting) |

BERTopic utilizes automatically computed probability thresholds to determine whether a document should be assigned to a topic or labeled as an outlier. Documents that cannot be confidently assigned to any topic based on this threshold are placed in the outlier topic labeled "-1" [17]. This threshold can be adjusted in the model parameter settings to reduce the number of outliers. In BERTopic 6, we used the same settings as in BERTopic 5. However, this time, we attempted to eliminate noise clusters using the 'reduce_outliers' function. This function takes documents and their related topics as input then reduce outlier documents and label them as non-linear topics. By embedding each outlier document, the most suitable topic embedding is determined using cosine similarity.

For example, in the tweet data in Table 4, the tweet "Jokowi achieves the target of reducing stunting rates, a comprehensive step towards progress." (*jokowi mencapai target penurunan angka stunting langkah terpadu semangat menuju nkri maju)* Using BERTopic 5 with datasets prep 1, 2, and 3, this tweet is identified as an outlier. After removing the outliers, this tweet is grouped into Topic 21, which has the following topic representation['Islam as a Solution to Life' (islamsolusikehidupan), 'Islam in Addressing Stunting' (islamatasistunting), 'Islam Solution to Life' (islamsolusikehidupan), ' Fate of the Generations' (nasib generasi), ' Muslim Women Speak' (muslimah bicara), ' Live (stream) by Muslim Women' (live muslimah)] This is done with data prep 1 and 2. In contrast, reducing outliers using prep 3, this tweet is grouped into Topic 1, with the following topic representation: ['food (pangan), resilience (ketahanan), granary (lumbung), national (nasional)']. Based on the results of qualitative observation data Prep 1 and 2, this tweet is grouped into topic 21, which we believe is not closely related even though it contains the word stunting. Still, this tweet is related to stunting reduction. This means that reducing noise has a negative impact on grouping each document because it forces a document into a topic that may not be related. Based on the grouping results of data prep 1 and 2, the tweet is grouped quite well because it contains the word stunting, but not for data prep 3, which groups outlier tweets into the food resilience topic. This is also consistent with the coherence values we obtained.

Data prep 1 produces the highest coherence value, followed by data prep 2 and prep 3, which is the smallest, so hashtag and mention information is quite useful for the model in grouping topics. Table 5 shows the coherence values generated by BERTopic 5 and 6 using datasets prep 1, 2, and 3. Based on Table 5, 'reduce_outliers' in BERTopic 6 reduces the coherence score generated by BERTopic 5. It is noted that BERTopic 5 and 6 have similar parameters, but in BERTopic 6, we perform outlier topic reduction. Therefore, this proves that removing outliers has a negative impact on topic grouping by reducing coherence between words within a topic, resulting in less coherent topics.

***Table 5:*** *Coherence values of BERTopic 5 and 6.*

| | **Prep 1** | **Prep 2** | **Prep 3** |
|---|---|---|---|
| **BERTopic5** | 0.7733 | 0.7612 | 0.7176 |
| **BERTopic6** | 0.7156 | 0.7095 | 0.7017 |

The number of topics generated by BERTopic, utilizing HDBSCAN, is considered valid as these methods are tailored to identify the optimal number of topics from the given data. However, it's important to acknowledge that the resulting number of topics might only sometimes align with the user's expectations or research goals. BERTopic employs a combination of UMAP for dimensionality reduction and HDBSCAN for document clustering to determine the number of topics. This process is designed to adapt to the data at hand. Nevertheless, the generated topics can vary based on data characteristics and model parameters. Manual evaluation and parameter adjustment are recommended to ensure the relevance and meaningfulness of the topics, such as modifying the min_cluster_size or tuning UMAP and HDBSCAN parameters to refine topic quality.

In contrast, traditional methods like k-means and agglomerative necessitate users to predefine the number of topics, potentially resulting in more interpretable topic structures. However, this approach may not accurately capture the data's inherent structure, leading to less coherent topics compared to those produced by HDBSCAN in BERTopic.HDBSCAN in BERTopic employs a hierarchical clustering strategy to automatically determine the number of topics, potentially yielding more coherent topic structures. This adaptive approach allows the model to identify optimal topic numbers based on the data, enhancing topic interpretability.

Meanwhile, for dimension reduction, UMAP is more suitable for high-dimensional data with complex relationships between data points, while PCA is more suitable for data with a linear structure. The choice between UMAP and PCA as dimension reduction techniques in BERTopic depends on the specific characteristics of the data and the research objectives.

In the context of applying BERTopic to Indonesian compared to English, the complexities of the Indonesian language need careful adjustment of numerous BERTopic parameters to improve topic extraction quality. Word embedding, dimension reduction, and clustering approaches are among the elements that contribute significantly to the model's adaptation to the linguistic peculiarities of Bahasa Indonesia.

**Word Embedding:** Because of the quick evolution of slang and flexible grammar, Indonesians must choose a suitable word embedding model. While English has well-established embeddings such as BERT and GloVe, Indonesian requires embeddings that have been carefully trained on local datasets to capture the intricacies of its changing vernacular and context-dependent word meanings. Tuning BERTopic to use such localized embeddings can greatly increase its capacity to understand and categorize Indonesian content.

**Dimension Reduction:** to cope with Indonesia's less organized grammar and high contextuality, dimension reduction approaches in BERTopic, such as UMAP or t-SNE, must be properly tuned. The best parameters for dimension reduction in Indonesian may differ from those in English due to the necessity to preserve the intricate relationships between words in the limited space, ensuring that the various meanings of words across different contexts are accurately represented.

**Clustering Method:** the clustering method utilized in BERTopic, such as HDBSCAN, must consider the many ways in which Indonesian is used on social media, where informal phrases and satire can affect the apparent topic coherence. The clustering method must be sensitive enough to distinguish between these nuances, which may necessitate different parameter values than those used for English in order to group related talks accurately.

**Parameter Tuning:** The parameters of BERTopic, such as the number of topics, threshold values for topic selection, and so on, must be fine-tuned to solve the special issues of the Indonesian language. For example, the criterion for subject size may need to be reduced to fit the more fragmented and diverse character of Indonesian social media debates.

Topic modeling in Indonesian is more challenging than in English due to its fluid syntax, frequently changing vocabulary and widespread use of humor and informal language on social media. These characteristics can result in large variances in word usage and meaning, necessitating a more complex approach to subject modeling. BERTopic's parameter tuning flexibility enables the customization required to successfully process and analyze Indonesian text, displaying versatility and high performance in languages with complex linguistic aspects.

## 5. CONCLUSION

In this study, we explored six topical parameter configurations within an Indonesian dataset focusing on biodiversity policy. We employed three distinct preprocessing scenarios: basic preprocessing, deletion of mentions, and elimination of both hashtags and mentions. Utilizing BERTopic, a sophisticated topic modeling technique, we analyzed the coherence and interpretability of the resulting topics.

BERTopic 1 utilized miniLM-L6-v2 for sentence embedding and HDBSCAN for clustering. Although coherent, we observed improved outcomes with Indonesian-specific sentence embeddings, as demonstrated by BERTopic 5 employing the 'Cahya Bert' embedding. BERTopic 2, utilizing KMeans and PCA for clustering, yielded a broader range of topics, even with universal sentence embeddings, with specific subject matter focusing on distinct entities. Interestingly, topics classified as outliers by HDBSCAN shared substantial information with non-outliers. BERTopic 3 employed Indonesian 'Firqaa'

sentence embedding combined with agglomerative clustering and UMAP, effectively preserving information. BERTopic 6 aimed at reducing outliers, like BERTopic 5, but often resulted in misclassification, complicating content interpretation. Our findings underscore the importance of selecting appropriate embeddings and clustering methods in topic modeling. While reducing outliers can refine topic categorization, it must be done cautiously to avoid distorting topic relevance and coherence.

Analysis of coherence scores across Bertopic models revealed significant influences of embedding size, language specificity, clustering techniques, and preprocessing methods. Models utilizing specific Indonesian embeddings consistently outperformed those using multilingual embeddings, highlighting the importance of linguistic specificity. Furthermore, data preparation method 3 consistently yielded better coherence scores, aligning with manual human interpretation.

In conclusion, our findings emphasize the multifaceted nature of coherence optimization in topic modeling. While various factors such as embedding size and specificity, clustering methods, and preprocessing techniques significantly impact coherence scores, manual interpretation remains crucial, particularly in scenarios of topic dispersion. Future research avenues may explore additional techniques for enhancing coherence in topic modeling, further advancing the field.

## ACKNOWLEDGEMENT

We want to express our sincerest gratitude to the Rumah Program Artificial Intelligence, Big Data, and Computational Technology for Biodiversity and Satellite Images 2023 for providing funding for the data collection in support of this project. This support was made possible by the Decree of the Head of BRIN number 1/III.6/HK/2023, and the project was successfully administered under the leadership of our head of research group, Mr. Mohammad Teduh Uliniansyah.

## AUTHOR CONTRIBUTIONS

Conceptualisation, Nuraisa Novia Hidayati; Methodology, Nuraisa Novia Hidayati, and Siti Shaleha; Formal analysis, Nuraisa Novia Hidayati, and Siti Shaleha; Investigation, Nuraisa Novia Hidayati, and Siti Shaleha; Data curation, Nuraisa Novia Hidayati and Siti Shaleha; Writing—original draft preparation, Nuraisa Novia Hidayati; Writing—original draft, Nuraisa Novia Hidayati and Siti Shaleha; Supervision and Review, Nuraisa Novia Hidayati. All authors have read and agreed to the published version of the manuscript.

## References

[1] K. von Rintelen, E. Arida, and C. Häuser, "A review of biodiversity-related issues and challenges in megadiverse Indonesia and other Southeast Asian countries," *Res. Ideas Outcomes*, vol. 3, Sep. 2017.

[2] S. Vasudeva Raju, B. Kumar Bolla, D. K. Nayak and J. Kh, "Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings," *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, Mumbai, India, pp. 1-6, 2022.

[3] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Front. Sociol.*, vol. 7, no. May, pp. 1–16, 2022.

[4] T. Ramamoorthy, V. Kulothungan, and B. Mappillairaju, "Topic modeling and social network analysis approach to explore diabetes discourse on Twitter in India.," *Front. Artif. Intell.*, vol. 7, p. 1329185, 2024.

[5] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia CIRP*, vol. 189, pp. 191–194, 2021.

[6] Z. A. Güven, B. Diri and T. Çakaloğlu, "Comparison Method for Emotion Detection of Twitter Users," *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, pp. 1-5, 2019.

[7] N. N. Hidayati and S. Rochimah, "Requirements traceability for detecting defects in agile software development," *EECCIS 2020 - 2020 10th Electr. Power, Electron. Commun. Control. Informatics Semin.*, pp. 248–253, 2020.

[8] N. N. Hidayati and A. Parlina, "Performance Comparison of Topic Modeling Algorithms on Indonesian Short Texts," in *Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications*, pp. 117–120, 2023.

[9] L. B. Hutama and D. Suhartono, "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," *Inform.*, vol. 46, no. 8, pp. 81–90, 2022.

[10] I. Scarpino, C. Zucco, R. Vallelunga, F. Luzza, and M. Cannataro, "Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration," *BioTech*, vol. 11, no. 3, Sep. 2022.

[11] G. Hristova and N. Netov, "Media Coverage and Public Perception of Distance Learning During the COVID-19 Pandemic: A Topic Modeling Approach Based on BERTopic," *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan, pp. 2259-2264, 2022.

[12] M. Asgari-Chenaghlu, M. R. Feizi-Derakhshi, L.

farzinvash, M. A. Balafar, and C. Motamed, "TopicBERT: A cognitive approach for topic detection from multimodal post stream using BERT and memory–graph," *Chaos, Solitons and Fractals*, vol. 151, p. 111274, 2021.

[13] M. de Groot, M. Aliannejadi, and M. R. Haas, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," 2022. [Online]. Available: `http://qwone.com/~jason/20Newsgroups/`.

[14] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," *Appl. Sci.*, vol. 13, no. 2, Jan. 2023.

[15] M. T. Uliniansyah *et al.*, "Twitter Dataset on Public Sentiments Towards Biodiversity Policy in Indonesia," *Data Br.*, vol. 52, p. 109890, 2024.

[16] S. Pebiana *et al.*, "Experimentation of Various Preprocessing Pipelines for Sentiment Analysis on Twitter Data about New Indonesia's Capital City Using SVM and CNN," *2022 25th Conf. Orient. COCOSDA Int. Comm. Co-ord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2022 - Proc.*, pp. 1–6, 2022.

[17] H. Axelborn and J. Berggren, "Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls," M.S. Thesis, UMEÅ University, Sweden, 2023.

[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022. [Online]. Available: `http://arxiv.org/abs/2203.05794`.

[19] O. Bulut, A. MacIntoshand and C. Walsh, "Using Lbl2Vec and BERTopic for Semi-Supervised Detec-tion of Professionalism Aspects in a Constructed-Response Situational Judgment Test," 2022. [Online]. Available: `osf.io/preprints/psyarxiv/n5fqe`.

[20] H. Lee, S. H. Lee, K. R. Lee, and J. H. Kim, "ESG Discourse Analysis Through BERTopic: Comparing News Articles and Academic Papers," *Comput. Mater. Contin.*, vol. 75, no. 3, pp. 6023–6037, 2023.

[21] M. Günther, L. Milliken, J. Geuter, G. Mastrapas, B. Wang, and H. Xiao, "Jina Embeddings: A Novel Set of High-Performance Sentence Embedding Models," 2023. [Online]. Available: `http://arxiv.org/abs/2307.11224`.

[22] Y. Yang *et al.*, "Multilingual universal sentence encoder for semantic retrieval," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 87–94, 2020.

[23] A. Perwira Joan Dwitama, D. Hatta Fudholi, and S. Hidayat, "Indonesian Hate Speech Detection Using Bidirectional Long Short-Term Memory (Bi-LSTM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 302–309, Mar. 2023.

[24] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," 2018, [Online]. Available: http://arxiv.org/abs/1802.03426.

[25] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," *WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min.*, pp. 399–408, 2015.

[26] M. S. Asyaky and R. Mandala, "Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP," *Proc. - 2021 8th Int. Conf. Adv. Informatics Concepts, Theory, Appl. ICAICTA 2021*, pp. 1–6, 2021.

**Nuraisa Novia Hidayati** her earned a Bachelor of Science in Computer Science from the Department of Information Systems at Sepuluh Nopember Institute of Technology (ITS) in 2012. She continued her studies at ITS, receiving a Master's degree in Computer Science from the Department of Informatics Engineering in 2021. She began her work as a software engineer, managing data in a variety of settings, including the telecommunications office, between 2012 and 2013. After that, she developed multiple human resources applications for the Agency for the Assessment and Application of Technology (BPPT) 2014-2021. Currently, she applies her knowledge as a researcher at the National Research and Innovation Agency (BRIN), focusing on Natural Language Processing (NLP), specifically text processing such as sentiment analysis, hoax detection, and topic modeling, to a variety of situations in Indonesia.

**Siti Shaleha** received a Bachelor of Applied Sciences in Electronics Engineering from the Electronic Engineering Polytechnic Institute of Surabaya, Indonesia, in 2018. She is currently working at the Research Center for Data and Information Sciences under the National Research and Innovation Agency of Indonesia. Her field of research includes artificial intelligence, natural language processing, and data sciences.