VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7204

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

EVALUATING CLUSTERING METHODS FOR SEMANTIC REPRESENTATION OF DISASTER NEWS USING BERT EMBEDDINGS AND HBDSCAN

Ariska Fitriyana Ningrum[1]*; Dannu Purwanto[1]; Abdel-Nasser Sharkawy[2]

Data Science[1]
Universitas Muhammadiyah Semarang, Semarang, Indonesia[1]
https://unimus.ac.id/[1]
ariskafitriyana@unimus.ac.id*, dannupurwanto@unimus.ac.id

Mechatronics Engineering, Mechanical Engineering Department, Faculty of Engineering[2]
Qena University, Qena, Egypt[2]
https://www.svu.edu.eg/en/[2]

Mechanical Engineering Department, College of Engineering[2]
Fahad Bin Sultan University, Tabuk, Saudi Arabia[2]
https://fbsu.edu.sa/[2]
abdelnassersharkawy@eng.svu.edu.eg

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Natural disasters that frequently occur in Indonesia demand a fast and accurate information monitoring and analysis system through online news sources. This study aims to identify topic patterns related to natural disasters in Indonesia using news articles from Detik.com through a semantic clustering approach. A total of 1,000 articles were collected, preprocessed, and represented using the Sentence-BERT (SBERT) model to capture contextual relationships between sentences. The vector representations were then clustered using three methods: K-Means, Agglomerative Hierarchical Clustering, and HDBSCAN. The performance of each method was evaluated using the Silhouette Score, Davies–Bouldin (DB) Index, and Calinski–Harabasz (CH) Index. The results show that HDBSCAN achieved the best performance with a Silhouette Score of 0.215, a DB Index of 1.557, and a CH Index of 18.102, outperforming Agglomerative (0.028, 3.945, 29.669) and K-Means (0.055, 3.678, 36.778). Moreover, the HDBSCAN model achieved the highest coherence score of 0.8669, indicating strong semantic consistency within clusters. Five coherent clusters emerged, representing major disaster themes: landslides, earthquakes, tornadoes, flash floods, and volcanic activity. The visualization of word clouds for each cluster reinforced the interpretation of these disaster topics. Overall, the combination of SBERT and HDBSCAN effectively groups news articles based on semantic similarity. These findings highlight the potential of Natural Language Processing (NLP) to enhance data-driven media monitoring, support early warning systems, and strengthen disaster communication and mitigation strategies in Indonesia*.

Keywords: *Natural Disasters News, Sentence BERT, Text Mining, Text Clustering.*

Intisari— *Bencana alam yang sering terjadi di Indonesia menuntut adanya sistem pemantauan dan analisis informasi yang cepat serta akurat melalui berita daring. Penelitian ini bertujuan untuk mengidentifikasi pola topik terkait bencana alam di Indonesia menggunakan berita dari situs Detik.com melalui pendekatan klasterisasi semantik. Sebanyak 1.000 artikel berita dikumpulkan, kemudian melalui tahap praproses dan direpresentasikan menggunakan model Sentence-BERT (SBERT) untuk menangkap hubungan kontekstual antar kalimat. Representasi vektor tersebut kemudian dikelompokkan menggunakan tiga metode klasterisasi, yaitu K-Means, Agglomerative Hierarchical Clustering, dan HDBSCAN. Kinerja masing-masing metode dievaluasi menggunakan tiga metrik, yaitu Silhouette Score, Davies–Bouldin (DB) Index, dan Calinski–*

Harabasz (CH) Index. Hasil evaluasi menunjukkan bahwa metode HDBSCAN memberikan kinerja terbaik dengan nilai Silhouette Score sebesar 0.215, DB Index 1.557, dan CH Index 18.102, mengungguli metode Agglomerative (0.028, 3.945, 29.669) dan K-Means (0.055, 3.678, 36.778). Selain itu, model HDBSCAN juga menghasilkan nilai coherence tertinggi sebesar 0.8669, yang menunjukkan konsistensi semantik yang kuat dalam setiap klaster. Dari hasil tersebut terbentuk lima klaster utama yang merepresentasikan tema bencana: tanah longsor, gempa bumi, angin puting beliung, banjir bandang, dan aktivitas vulkanik. Visualisasi word cloud pada masing-masing klaster memperkuat interpretasi tema bencana tersebut. Kombinasi antara SBERT dan HDBSCAN terbukti efektif dalam mengelompokkan berita berdasarkan kesamaan makna, serta berpotensi meningkatkan sistem pemantauan media berbasis data, mendukung sistem peringatan dini, dan memperkuat strategi mitigasi bencana di Indonesia.

***Kata Kunci**: Berita Bencana Alam, Klasterisasi Teks, Penambangan Teks, Sentence BERT.*

INTRODUCTION

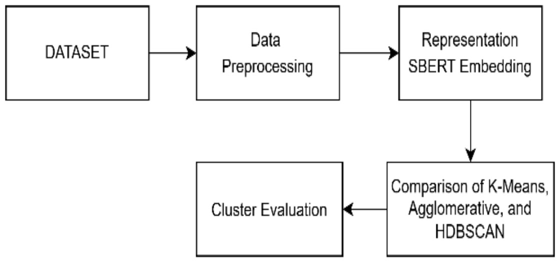
Natural disasters are a global issue that not only cause physical and social damage but also generate massive volumes of textual data from mass media, social media, and disaster reporting systems [1]. This information is often unstructured, dispersed, and rapidly expanding, making it difficult to monitor and support real-time decision-making. In today's digital era, disaster-related information is widely disseminated through various online platforms, creating an urgent need for efficient information management and analysis to support disaster mitigation and rapid response efforts [2], [3]. To address these challenges, computational approaches from the field of Natural Language Processing (NLP) have emerged as powerful tools for automatically organizing and interpreting disaster-related text data. Among these, clustering plays a vital role in grouping similar pieces of information and uncovering hidden semantic patterns within large volumes of news data [1].

The integration of semantic representation models such as Sentence-BERT (SBERT) with modern clustering algorithms offers a significant opportunity to extract meaningful insights from complex and dynamic text collections, helping both stakeholders and the public better understand disaster situations [4] [5]. Text clustering, as a core technique in data mining and information retrieval, aims to organize semantically similar documents into coherent groups. Various approaches have been proposed—from classical methods such as K-Means and DBSCAN to advanced density-based algorithms like Hierarchical Density-Based Clustering [6] and HDBSCAN [7]. The effectiveness of these methods, however, strongly depends on the quality of text feature representation. Advances in semantic embedding models such as SBERT enable text to be mapped into high-dimensional vector spaces that preserve sentence-level meaning. Compared to traditional methods like TF-IDF or Word2Vec, SBERT provides deeper contextual understanding [8]. This is particularly important for disaster-related news, which often contains synonyms, idiomatic expressions, and local linguistic variations that can affect clustering performance. Several previous studies have examined clustering and embedding methods for disaster-related text Chanda [9] demonstrated the effectiveness of BERT in predicting disasters from Twitter data, while J. Li and B. Li. [10] and Mohammed Alsuhaibani [11] emphasized the value of semantic approaches for topic and event detection in online news. Other studies [12], [13] have evaluated BERT embeddings with clustering algorithms such as DBSCAN and HDBSCAN, yet few have explicitly applied and compared SBERT-based semantic representations within the disaster news domain. Moreover, limited research has quantitatively assessed clustering quality using metrics such as the Silhouette Score [14] and qualitatively evaluated topic coherence. Kapellas et al [15] and Alasalı and Ortakcı [16]highlighted the importance of selecting clustering techniques suited to dynamic and unstructured news data.

Building upon this gap, this study explicitly hypothesizes that combining SBERT semantic embeddings with density-based clustering methods such as HDBSCAN yields more coherent and interpretable topic clusters compared to traditional algorithms like K-Means or Agglomerative Clustering. Therefore, the main objective of this research is to evaluate and compare various clustering algorithms on SBERT based semantic representations of Indonesian natural disaster news. The study aims to identify which method provides the most effective grouping both technically based on quantitative performance metrics and semantically based on interpretability and coherence. Ultimately, this research contributes to the field of text mining and disaster informatics by proposing an evidence-based framework for automatic disaster topic detection, which can enhance data-driven media monitoring, early warning systems, and disaster response strategies in Indonesia.

MATERIALS AND METHODS

This study began with the collection of 1,000 natural disaster news articles from *detik.com*, including titles, content, dates, and locations. The data was preprocessed through text cleaning, stopword removal, normalization, and merging into a single *clean_text* column. Semantic representation was performed using the SBERT *all-MiniLM-L6-v2* model to generate embedding vectors for each article. SBERT was selected over traditional approaches such as TF-IDF or word2vec due to its superior ability to capture contextual and semantic similarity between sentences, making it particularly effective for clustering semantically related news texts. Subsequently, three clustering algorithms—K-Means, Agglomerative, and HDBSCAN—were applied to group articles based on semantic similarity. Cluster evaluation was conducted using internal metrics (Silhouette, Davies-Bouldin, Calinski-Harabasz) as well as visual explorations such as WordCloud, time distribution, and disaster type. The results of each method were compared to determine the most optimal clustering algorithm for SBERT-based representations. Research stages are shown in Figure 1.


DATASET
Data Preprocessing
Representation SBERT Embedding
Comparison of K-Means, Agglomerative, and HDBSCAN
Cluster Evaluation

Source : (Research Results, 2025)
Figure 1. Research Stages

A. Dataset

The data used in this study is a collection of news articles about natural disasters obtained from the Detik.com website. Data collection was carried out using web scraping techniques with Python libraries, namely requests and Beautiful Soup, by searching for articles containing keywords such as "natural disasters." The scraping process was carried out on 100 Detik.com web pages and produced 1,000 articles containing news related to natural disasters. All articles were collected and stored in CSV format, with data attributes including headline, publication date, link, and article content. The data obtained is presented in Table 1.

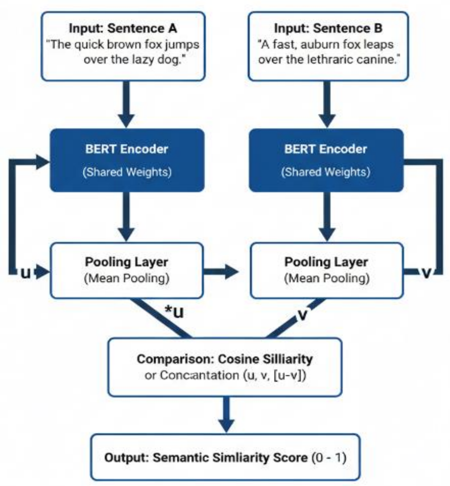
Table 1. Dataset Features

Feature	Description
Headline	The main of an article or news item that summarizes the core information concisely and attracts the reader's attention
Date	The publication date of the article or news item, indicating when the information was published
Link	The URL address that directs the reader to the source article or news item on the relevant website
Content	The full text or excerpt from the article/news item containing the details of the information presented in the writing.

Source : (Research Results, 2025)

B. Sentence Bidirectional Encoder Representation from Transformer (S-BERT)

BERT is a pre-trained language model based on a transformer encoder that captures the contextual meaning of words. It learns word representations by considering surrounding words, enabling effective handling of various NLP tasks [13]. However, directly averaging BERT's word embeddings to represent sentences often fails to capture their overall semantic meaning, which is crucial for tasks such as clustering and semantic similarity [14], [15], [16]. As a solution, Mohammed Alsuhaibani [11] roposed a streamlined alternative to SBERT by directly exploiting BERT's internal layers through a max pooling strategy, avoiding complex Siamese architectures. The study demonstrates that extracting features from the 7th layer provides a computationally efficient yet robust approach for entailment detection on the SNLI dataset. An illustration of the Sentence-BERT architecture is shown in Figure 2.


Input: Sentence A
"The quick brown fox jumps over the lazy dog."
Input: Sentence B
"A fast, auburn fox leaps over the lethraric canine."
BERT Encoder
(Shared Weights)
BERT Encoder
(Shared Weights)
u
Pooling Layer
(Mean Pooling)
Pooling Layer
(Mean Pooling)
v
*u
v
Comparison: Cosine Silliarity
or Concantation (u, v, [u-v])
Output: Semantic Simliarity Score (0 - 1)

Source : (Research Results, 2025)
Figure 2. Sentence BERT architecture

As illustrated in Figure 2, Sentence A and Sentence B are encoded by identical BERT encoders to generate contextualized token embeddings. A pooling layer converts these embeddings into fixed-length vectors $(u \, and \, v)$. which represent the semantic meaning of each sentence. The similarity between sentences is then computed using cosine similarity.

$$C_V = \frac{1}{\binom{N}{2}} \sum_{i<j} \cos(\vec{v}_i, \vec{v}_j) \quad (1)$$

where $\vec{v_i}$and $\vec{v_j}$are the term vectors derived from a sliding window co-occurrence model. Higher $C_V$values indicate greater semantic consistency within a cluster, thereby supporting the claim of semantic validity in the topic structure.

C. Clustering

Agglomerative Hierarchical Clustering s a *bottom-up unsupervised learning* method that progressively merges the most similar documents until a hierarchical cluster structure is formed. In news text analysis, AHC is used to group articles based on semantic similarity to identify dominant themes or issues without requiring labeled data. The theoretical stages include text preprocessing (tokenization, stopword removal, lemmatization), vector representation using *TF-IDF* or *Sentence-BERT*, and similarity computation between documents using cosine similarity [17].

$$Sim(A, B) = \frac{A.B}{\|A\|\|B\|} \quad (2)$$

The distance between clusters is then calculated using a *linkage function*, such as the Ward method, which minimizes within-cluster variance:

$$D(A, B) = \frac{|A| \, .|B|}{|A|+|B|} \|\bar{x}_A - \bar{x}_B\|^2 \quad (3)$$

The merging process is visualized through a dendrogram, while cluster quality is assessed using internal validation metrics such as the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. This approach effectively uncovers topic structures and semantic relationships across news articles and other textual data. Recent studies supporting this methodology include *Hierarchical Level-Wise News Article Clustering via Multilingual Matryoshka Embeddings* (2025), which applies agglomerative hierarchical clustering with multilingual embeddings to capture semantic similarity at different topic levels; *Text Mining Customer Feedback: An Agglomerative Clustering Approach* (2025), which integrates text preprocessing, TF-IDF, PCA, and AHC with Silhouette-based evaluation on real-world textual data; and *Semantic Deep Embedded Clustering* (2025), which leverages transformer-based embeddings to significantly enhance semantic understanding in large-scale text clustering tasks. [18], [19], [20]

K-means Clustering is an unsupervised learning algorithm that clusters data into K clusters by minimizing the distance between data points and cluster centroids. The goal is to minimize the objective function[21]:

$$J = \sum_{j=1}^{K} \sum_{x_i \in S_j} \left\| x_i - C_j \right\|^2 \quad (4)$$

with:
C_j = cluster centroid *to j*
x_j = data vector (in thi study, sentence embedding from Sentence-BERT)
S_j = set of data included in cluster to$-j$
In the context of natural disaster news, the vector embeddings generated by Sentence-BERT represent the semantic meaning of each news item. K-Means then clusters the news items based on topic similarity, such as floods, earthquakes, or forest fires [22]. The optimal K value can be determined using internal evaluation metrics such as Silhouette Score, Davies–Bouldin Index, or Calinski–Harabasz Index to ensure the quality of the clusters formed.

HDBSCAN (Hierarchical DBSCAN)

HDBSCAN consists of four main stages in forming clusters. The first step estimates the density of each point by calculating the distance to its $k-$ *nearest neighbor*, referred to as the core distance for parameter k. To distinguish low density points or noise, HDBSCAN introduces a new distance metric called the mutual reachability distance, defined as [13].

$$d_{mreach-k}(a, b) = \max\{core_k(a), core_k(b), d(a, b)\} \quad (5)$$

Where $d(a, b)$ is the original distance between points a and b, and $core_k$ represents the density estimate for parameter k. This mutual reachability distance is then used to construct a Minimum Spanning Tree (MST) that connects dense regions in the dataset. The third and most critical step of HDBSCAN is tree pruning, which involves comparing the number of points within each branch to the predefined minimum cluster size. The final step is to compute the cluster stability from the pruned tree using the following equation [23]

$$\sum_{p \epsilon cluster} (\lambda_p - \lambda_{birth}) \quad (6)$$

Here, λ_{birth}denotes the threshold value at which a cluster is formed, while $\lambda_p$indicates the threshold at which a point leaves the cluster. The stability of each cluster determines whether it is included in the final clustering result. HDBSCAN also includes several hyperparameters that can be tuned based on the dataset characteristics [24]. The most important one is the minimum cluster size, which specifies the smallest number of points required to form a cluster. If a group of points does not meet this minimum, those points are treated as noise.

RESULTS AND DISCUSSION

The analysis of natural disaster news from detik.com was conducted to explore temporal patterns, semantic representations, and topic classifications of disasters in Indonesia. A total of 1,000 articles were scraped from 100 web pages and stored in a structured dataset. After data collection, a series of preprocessing steps were applied to the “content” column to ensure text quality before analysis. The steps included: (1) converting all text to lowercase, (2) removing numbers, punctuation, and extra spaces, (3) removing stopwords, and (4) applying lemmatization or stemming to normalize words to their base forms. These steps produced clean and consistent text data for subsequent SBERT-based embedding and clustering.

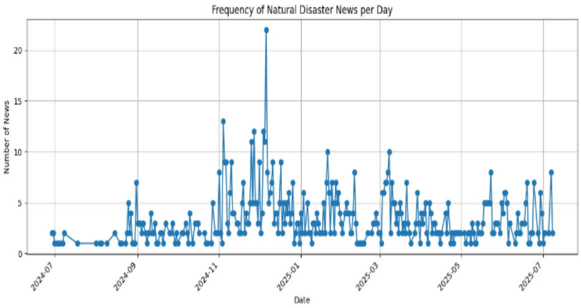
In addition, the dataset was examined for potential data imbalance by analyzing the distribution of articles across disaster types and time periods to ensure adequate representation. This chapter presents visualizations of daily news frequencies, clustering outcomes, and the main emerging topics. The findings provide insights into media focus on disaster issues and highlight the potential of NLP for early warning and disaster management.

Table 2 . Dataset After Preprocessing (Dataset in Indonesian)

Headline	Content	Date	Time
450 KK Terdampak Banjir di Lombok Barat, Dua Ru...	Banjir bandang melanda sejumlah wilayah di Kab...	08-Jul-25	08.40
Gunung Lewotobi Laki-laki di Lewotolok Me...	Dua gunung di Nusa Tenggara Timur (NTT) meletu...	08-Jul-25	07.09
Gunung Lewotobi Meletus Lagi Malam Ini, Warga...	Gunung Lewotobi Laki-laki di Kabupaten Flores ...	07-Jul-25	19.51

Source : (Research Results, 2025)

The results of preprocessing (Table 2) on the content variable show that the news content has been cleaned of irrelevant elements such as HTML tags, special punctuation marks, and other potentially insignificant words, leaving only the main text containing the core information of each disaster news item. This cleaned text is then better prepared for further processing, such as word frequency analysis, topic classification, or text-based machine learning model creation. Additionally, this preprocessing helps reduce noise and improve the accuracy of text analysis results by avoiding interference from meaningless characters or symbols. Meanwhile, for the date variable, the preprocessing process separates the combined date and time information into two separate variables: date and time. This process facilitates exploration of the distribution of news over time. One form of initial analysis that can be performed is to look at the frequency of natural disaster news coverage per day, which is visualized in the following graph. This analysis provides an overview of the intensity patterns of disaster news coverage, which can be used to identify periods with the highest news coverage activity and evaluate the temporal trends of disaster issues reported by online media.



Source : (Research Results, 2025)

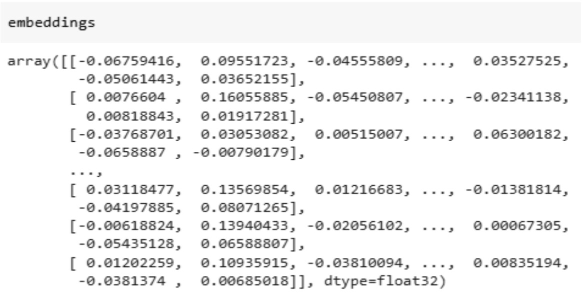
Figure 3. Daily frequency trend of natural disaster news coverage

The graph of daily news frequency (Figure 3) trends for natural disasters reveals a clear temporal pattern throughout the observation period. A substantial surge in reporting began around November 2024, peaking in late December 2024, when daily news coverage exceeded 20 articles per day. Prior to this peak, coverage intensity remained relatively low and stable, typically between 1 to 3 articles per day. This sharp escalation suggests the occurrence of major disaster events or multiple disasters in close succession, which drew heightened media attention. Following January 2025, the frequency of disaster-related news gradually declined but continued to fluctuate, maintaining moderate attention with fewer than 10 reports per day. This pattern reflects not only

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

seasonal factors, such as Indonesia's rainy season that often coincides with floods and landslides, but also periodic geological risks, including volcanic or seismic activity that influences media focus. The persistent, albeit fluctuating, coverage indicates that natural disasters remain a continuous public concern in Indonesia's news cycle.

These findings are consistent with the observations of Tounsi et al. (2023), who found that media attention to disaster events typically spikes during periods of high-impact hazards, followed by gradual attenuation once immediate crises pass. Similarly, Wang et al. (2024) noted that the temporal distribution of disaster-related reports often aligns with environmental and climatic cycles, reflecting how information dissemination mirrors real-world disaster seasonality [25], [26]. In contrast to prior studies that primarily captured global disaster reporting trends, the current results provide a localized temporal insight, emphasizing the seasonal and event-driven nature of disaster communication in Indonesian online media. After the clean_text data has been cleaned, the text representation is transformed into a numerical vector using the Sentence-BERT (SBERT) model. SBERT is a transformer-based embedding model specifically designed to generate meaningful sentence representations in a fixed vector dimension. The primary purpose of this representation is to capture the semantics of each news article, so that news articles with similar meanings will have vectors that are relatively close in the vector space. This is crucial as an initial step before the clustering process is performed, as the quality of the representation will significantly influence the clustering results. Figure 4 presents the results of text embedding using SBERT.



```
embeddings

array([[-0.06759416,  0.09551723, -0.04555809, ...,  0.03527525,
        -0.05061443,  0.03652155],
       [ 0.0076604 ,  0.16055885, -0.05450807, ..., -0.02341138,
         0.00818843,  0.01917281],
       [-0.03768701,  0.03053082,  0.00515007, ...,  0.06300182,
        -0.0658887 , -0.00790179],
       ...,
       [ 0.03118477,  0.13569854,  0.01216683, ..., -0.01381814,
        -0.04197885,  0.08071265],
       [-0.00618824,  0.13940433, -0.02056102, ...,  0.00067305,
        -0.05435128,  0.06588807],
       [ 0.01202259,  0.10935915, -0.03810094, ...,  0.00835194,
        -0.0381374 ,  0.00685018]], dtype=float32)
```

Source : (Research Results, 2025)

Figure 4. Result of text embedding using BERT

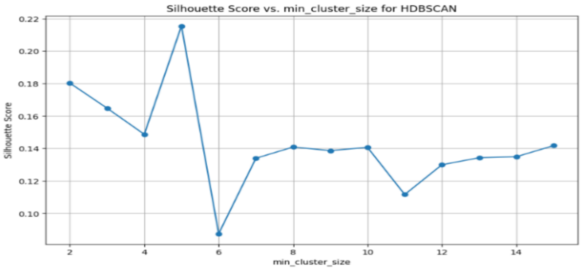
After the text representation process is carried out using Sentence-BERT (SBERT), the next step is to group the data into several clusters based on the similarity of meaning between sentences. Clustering is an important stage in exploring hidden topics because it allows us to identify structures or patterns in large text collections. In this study, three clustering approaches were employed K-Means, Agglomerative Clustering, and HDBSCAN to compare their performance and the interpretability of their results. Table 3 presents the comparison results of the accuracy of the three clustering methods.

Table 3. Clustering Evaluation Results

Method	Sillhoutte Score	DB Index	CH Index
Agglomerative	0.028	3.945	29.669
K-Means	0.055	3.678	36.778
HDBSCAN	0.215	1.557	18.102

Source : (Research Results, 2025)

Based on the evaluation of the three clustering methods Agglomerative, K-Means, and HDBSCAN it can be concluded that HDBSCAN achieved the best performance in clustering the SBERT-embedded text data. This conclusion is supported by its highest Silhouette Score of 0.215, which, although moderate in absolute value, indicates that the clusters are relatively well-separated and internally cohesive compared to other methods. In clustering text embeddings, such a score often reflects meaningful but not perfectly distinct semantic groupings due to the inherent overlap of linguistic concepts. Furthermore, HDBSCAN recorded the lowest Davies–Bouldin Index (1.557), suggesting that the formed clusters are compact and exhibit minimal overlap—an important quality for high-dimensional textual data. While the Calinski–Harabasz (CH) Index of HDBSCAN is lower than that of K-Means, this difference may arise from HDBSCAN's ability to handle noise and variable-density clusters, which the CH Index does not fully capture. Overall, the combination of a relatively high Silhouette Score and a low DBI highlights HDBSCAN's effectiveness in producing semantically coherent and well-structured clusters. In contrast, the Agglomerative method produced the weakest performance, as indicated by its low Silhouette and CH Index scores and high DBI value, which collectively reflect poorly separated and internally inconsistent clusters..



Source : (Research Results, 2025)

Figure 5. Silhouette Score

Figure 5 shows the Silhouette Score graph against the min_cluster_size value in the HDBSCAN algorithm.

1. The highest Silhouette Score (0.217) occurs when min_cluster_size = 5, indicating that at a minimum cluster size of 5, the separation and density between clusters are optimal.
2. A min_cluster_size value of 6 yields the lowest score (0.089), indicating that clusters formed at this size are less effective in terms of cohesion and separation.
3. After min_cluster_size = 6, the score begins to rise again but remains within a lower range compared to its peak at the value of 5.

So that cluster 5 is the most optimal for clusteirng this data sccording to the Sillhouette Score. This claue provides the best balance between the number of cluster members and the quality of separation between clusters. After the text representation process using Sentence-BERT (SBERT) adn the evaluation of performance of various clustering methods, HDBSCAN was selected as the best algorithm based on the highest Silhouette Score Value (0.215) and the lowest DB Index (1.557).

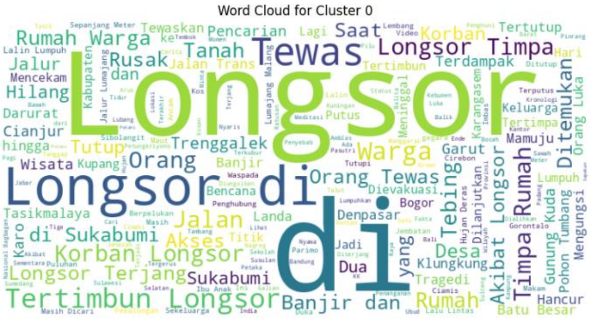
In addition, the clusters generated by HDBSCAN also achieved the highest Coherence Score (0.8669), indicating that the terms within each cluster are sematically consistent and represent coherent topics. This suggests that the clustering results are not only well separated geometrically (as shown by the Sillhoutte and DBI metrics) but also meaningful from a semantic perspective. Next, an analysis was conducted on the distribution of members in each cluster produced by HDBSCAN to further understand the data grouping patterns.

Table 4. Number of Cluster Members

Cluster	Member of Cluster
0	179
1	63
2	144
3	195
4	419

Source : (Research Results, 2025)

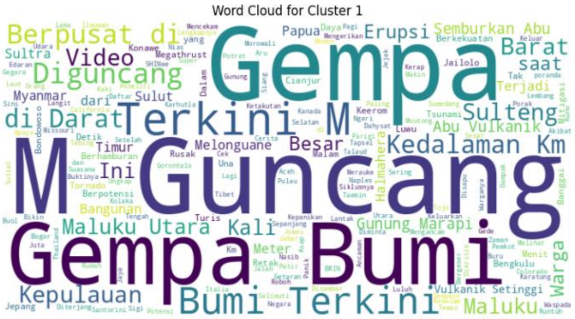
The clustering results with HDBSCAN, presented in Table 4, produced five clusters with varying numbers of members, namely 179, 63, 144, 195, and 419 members. This variation in numbers shows that HDBSCAN is able to adaptively identify different data densities without the need to determine the number of clusters at the outset. The largest cluster is Cluster 4 with 419 members, indicating the presence of a fairly dense and consistent data group, while Cluster 1 is the smallest with 63 members, possibly reflecting more scattered data or containing specific topics. The relatively balanced and not overly small cluster composition suggests that the semantic structure of the embedding results has been captured effectively, and the clustering process has been optimized. To understand the characteristics of each cluster, visualization was performed using a word cloud to identify the most prominent words in each group as a representation of the dominant topics contained therein.



Source : (Research Results, 2025)

Figure 6. WordCloud for Cluster 0 about Landslide Events

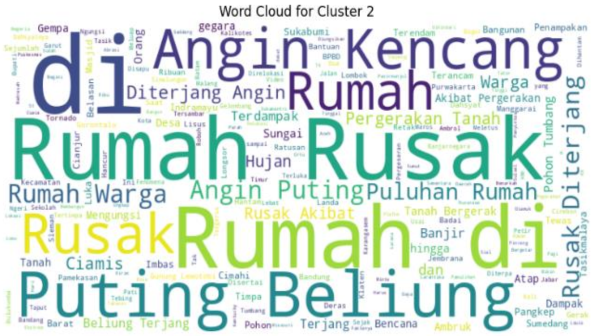
Based on the word cloud visualization for Cluster 0 (Figure 6), the word *"Longsor"* appears as the most dominant term, signifying that this cluster centers on landslide events. The term *"longsor"* appears with a normalized frequency of 6.8%, dominating the cluster. Supporting words such as *"tewas"* (4.3%) and *"tertimbun"* (3.9%) indicate the emphasis on casualties and damage. The co-occurrence of location names (e.g., *Sukabumi*, *Cianjur*) confirms regional event reporting. This cluster reflects strong event-focused narratives similar to Fauzi et al. (2021), but with higher semantic granularity emphasizing the response and rescue phase, as shown by action verbs (*evakuasi*, *ditemukan*)[27], [28].



Source : (Research Results, 2025)

Figure 7. WordCloud For Cluster 1 about Earthquakes and Volcanic Activity

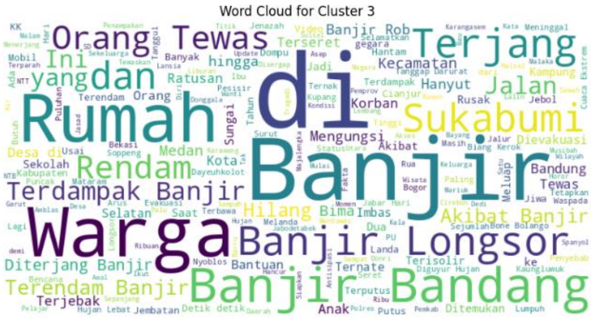
In Cluster 1 (Figure 7), dominant words such as *"Gempa," "Bumi," "Guncang,"* and *"Terkini"* indicate that the cluster discusses earthquake phenomena. Dominated by *"gempa"* (7.1%) and *"guncang"* (4.8%), this cluster links seismic activities with regional mentions such as *Maluku* and *Sulawesi*. The TF-IDF weight for *"episentrum"* (0.034) suggests a technical orientation in the discourse, integrating both pu.blic reactions and scientific parameters [29] .



Source : (Research Results, 2025)

Figure 8. WordCloud For Cluster 2

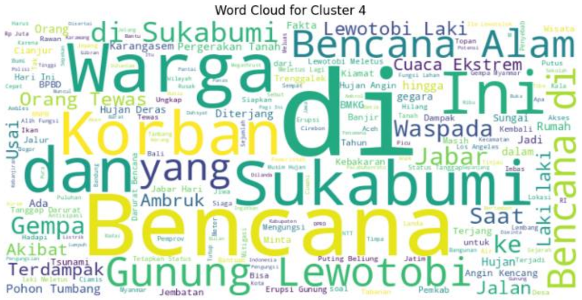
The Cluster 2 word cloud shown in Figure 8 highlights topics related to strong winds or tornadoes, characterized by terms such as "damaged," "struck," and "residents". Key terms like "angin" (5.9%) and "rusak" (4.2%) reveal a focus on wind-induced damage. The inclusion of place names (Ciamis, Cimahi, Tasikmalaya) shows localized incident reporting. Compared with Wang et al. (2024), the frequency distribution suggests that damage-oriented narratives (45% of top terms) dominate over general event descriptions, highlighting social and infrastructural impacts [26].



Source : (Research Results, 2025)

Figure 9. WordCloud For Cluster 3 about Strong Winds / Tornadoes

Cluster 3 as shown in Figure 9 captures discussions on flooding disasters, as reflected by dominant terms such as *"flood," "residents," "houses," "affected,"* and *"submerged."* Words like *"bandang"* and *"terjang"* denote the severity of the flood, while *"killed," "missing,"* and *"evacuated"* reveal human and infrastructural impacts. With *"banjir"* showing the highest frequency (8.3%), this cluster demonstrates concentrated discussions on flood events and their impacts. Terms like *"terendam"* (4.5%) and *"hanyut"* (3.8%) strengthen the focus on severity. The flood narrative integrates both the event phase and post-event response, consistent with the work of Tounsi et al [25].



Source : (Research Results, 2025)

Figure 10. WordCloud For Cluster 4 about Lewotobi Volcanic Eruption

Finally, Cluster 4 (Figure 10) pertains to the Lewotobi volcanic eruption in East Flores, Nusa Tenggara Timur. Terms such as *"mountain," "Lewotobi," "eruption,"* and *"vulkanik"* indicate volcanic activity, while *"residents," "victims," "affected," "evacuation,"* and *"tanggap darurat"* illustrate the emergency response and mitigation phases. The terms *"Lewotobi"* (6.5%) and *"erupsi"* (5.1%) dominate, highlighting the event's specificity. Words such as *"evakuasi"* (3.9%) and *"tanggap darurat"* (2.7%) demonstrate the transition from hazard occurrence to mitigation and emergency response. This dual-phase theme (event + response) aligns with Indrani et al. (2023).

Overall, while previous studies (Tounsi et al., 2023; Fauzi et al., 2021; Huang et al., 2022; Wang et al., 2024] have demonstrated that clustering can effectively group disaster-related text by event type, this study extends the discussion by capturing multi-level thematic differentiation from event type to response actions and regional specificity [25], [26], [28], [30]. Such results highlight the potential of clustering models in semantic representation of disaster narratives, particularly within the Indonesian linguistic and geographic context, providing new insights for information extraction and disaster management communication.

CONCLUSION

This study demonstrated the effectiveness of combining Sentence-BERT (SBERT) embeddings with the HDBSCAN clustering algorithm in identifying semantic patterns and topic structures within disaster-related news articles from Detik.com. The evaluation results revealed that HDBSCAN achieved the best overall performance, with the highest Silhouette Score (0.215), the lowest Davies-Bouldin Index (1.557), and a high Coherence Score (0.8669). These metrics indicate that the clusters produced were well-separated and semantically coherent, confirming the suitability of the SBERT–HDBSCAN framework for disaster news analysis. Nonetheless, several limitations should be acknowledged. The dataset was limited to news articles from a single source (Detik.com), which may introduce source bias due to specific editorial styles, linguistic patterns, and reporting priorities unique to that platform. Such bias could lead to an overrepresentation of particular disaster narratives or regions, thereby reducing the model's ability to generalize to broader media ecosystems. For instance, disaster coverage in other outlets (e.g., Kompas, CNN Indonesia) or on social media platforms may employ different vocabulary, sentiment, or framing, potentially resulting in divergent clustering outcomes. Moreover, the model did not explicitly address the imbalanced distribution of disaster types and temporal variations in reporting frequency, which might influence topic salience and inter-cluster relationships. Despite these constraints, the findings underscore the practical potential of this approach for real-world disaster management applications. Automated clustering of disaster-related news can enhance early warning systems, support rapid situational awareness, and assist authorities in monitoring media coverage and public discourse during crisis events. Furthermore, this framework can be integrated into disaster information dashboards to improve data-driven decision-making. Future research should extend this work by incorporating multi-source datasets from various news platforms and social media to improve representativeness and minimize bias. In addition, applying temporal and geospatial modeling would enable a deeper understanding of how disaster topics evolve over time and across regions. Integrating sentiment analysis or event extraction techniques could also enrich interpretability and operational value, contributing to the development of comprehensive, real-time disaster.

REFERENCE

[1] X. Chen, "Monitoring of Public Opinion on Typhoon Disaster Using Improved Clustering Model Based on Single-Pass Approach," *Sage Open*, vol. 13, no. 3, Jul. 2023, doi: 10.1177/21582440231200098.

[2] R. Mena, "Advancing 'no natural disasters' with care: risks and strategies to address disasters as political phenomena in conflict zones," *Disaster Prevention and Management: An International Journal*, vol. 32, no. 6, pp. 14–28, 2023, doi: 10.1108/DPM-08-2023-0197.

[3] F. Sufi and M. Alsulami, "AI-Driven Global Disaster Intelligence from News Media," *Mathematics*, vol. 13, no. 7, Apr. 2025, doi: 10.3390/math13071083.

[4] L. Wang *et al.*, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," Feb. 2024, doi: https://doi.org/10.48550/arXiv.2212.03533.

[5] A. O. Alharm and S. Naim, "Enhancing Natural Disaster Response: A Deep Learning Approach to Disaster Sentiment Analysis using BERT and LSTM," 4755638, 2024. doi: 10.2139/ssrn.4755638.

[6] M. S. Asyaky and R. Mandala, "Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP," in *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAICTA53211.2021.9640285.

[7] G. Stewart and M. Al-Khassaweneh, "An Implementation of the HDBSCAN* Clustering Algorithm," *Applied Sciences (Switzerland)*, vol. 12, no. 5, Mar. 2022, doi: 10.3390/app12052405.

[8] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[9] A. K. Chanda, "Efficacy of BERT embeddings on predicting disaster from Twitter data," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.10698

[10] J. Li and B. Li, "Topic Mining of Civil Aviation Supervision Texts Based on BERTopic Model," in *Proceedings of the 2025 5th International Conference on Internet of Things and Machine Learning, IoTML 2025*, Association for Computing Machinery, Inc, Aug. 2025, pp. 176–183. doi: 10.1145/3749566.3749605.

[11] M. Alsuhaibani, "Deep Learning-based Sentence Embeddings using BERT for Textual Entailment," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, p. 2023, 2023,

doi: https://doi.org/10.14569/IJACSA.2023.01408108.

[12] A. Subakti, H. Murfi, and N. Hariadi, “The performance of BERT as data representation of text clustering,” *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00564-9.

[13] Y. Zhang, Z. Chen, X. Zheng, N. Chen, and Y. Wang, “Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data,” *J. Hydrol. (Amst).*, vol. 603, Dec. 2021, doi: 10.1016/j.jhydrol.2021.127053.

[14] I. Firman Ashari, E. Dwi Nugroho, R. Baraku, I. N. Yanda, and R. Liwardana, “Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta,” 2023. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[15] N. Kapellas and S. Kapidakis, “Event Detection in News Articles: A Hybrid Approach Combining Topic Modeling, Clustering, and Named Entity Recognition,” in *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings*, Science and Technology Publications, Lda, 2023, pp. 272–279. doi: 10.5220/0012234300003598.

[16] T. Alasali and Y. Ortakaci, “Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications,” *Computer Science*, vol. 9, pp. 32–50, Mar. 2024, doi: 10.53070/bbd.1421527.

[17] A. Polimeno, M. Reuver, S. Vrijenhoek, and A. Fokkens, “Improving and Evaluating the Detection of Fragmentation in News Recommendations with the Clustering of News Story Chains,” Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.06192

[18] M. W. U. Rahman, R. Nevarez, L. T. Mim, and S. Hariri, “SDEC: Semantic Deep Embedded Clustering,” Aug. 2025, [Online]. Available: http://arxiv.org/abs/2508.15823

[19] L. Muthoharoh, “Text Mining Customer Feedback: An Agglomerative Clustering Approach to Service Optimization,” *International Journal of Electronics and Communications Systems*, vol. 5, no. 1, pp. 31–51, Jun. 2025, doi: 10.24042/ijecs.v5i1.27188.

[20] H. W. A. Hanley and Z. Durumeric, “Hierarchical Level-Wise News Article Clustering via Multilingual Matryoshka Embeddings,” May 2025, [Online]. Available: http://arxiv.org/abs/2506.00277

[21] M. Fuchs and W. Höpken, “Clustering: Hierarchical, k-Means, DBSCAN,” in *Tourism on the Verge*, vol. Part F1051, Springer Nature, 2022, pp. 129–149. doi: 10.1007/978-3-030-88389-8_8.

[22] R. Kusumaningrum, S. F. Khoerunnisa, K. Khadijah, and M. Syafrudin, “Exploring Community Awareness of Mangrove Ecosystem Preservation through Sentence-BERT and K-Means Clustering,” *Information (Switzerland)*, vol. 15, no. 3, Mar. 2024, doi: 10.3390/info15030165.

[23] M. S. Asyaky and R. Mandala, “Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP,” in *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAICTA53211.2021.9640285.

[24] M. S. Asyaky and R. Mandala, “Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP,” in *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAICTA53211.2021.9640285.

[25] A. Tounsi and M. Temimi, “A systematic review of natural language processing applications for hydrometeorological hazards assessment,” Apr. 01, 2023, *Springer Science and Business Media B.V.* doi: 10.1007/s11069-023-05842-0.

[26] Z. Wang, X. Shi, H. Yang, B. Yu, and Y. Cai, “Automatic Extraction and Cluster Analysis of Natural Disaster Metadata Based on the Unified Metadata Framework,” *ISPRS Int. J. Geoinf.*, vol. 13, no. 6, Jun. 2024, doi: 10.3390/ijgi13060201.

[27] A. D. P. Ariyanto, D. Purwitasari, C. Fatichah, S. D. Ravana, Andrian, and A. A. Y. Parwata, “Transformer-Based Semantic Role Labeling for Crisis Events Using Semi-Supervised Learning on Low-Resource Language Twitter Texts,” *IEEE Access*, vol. 13, pp. 158938–158966, 2025, doi: 10.1109/ACCESS.2025.3604068.

[28] Mustakim, Muhammad Zakiy Fauzi, Mustafa, Assyari Abdullah, and Rohayati, “Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms ,” *J. Phys. Conf. Ser.*, vol. 1, 2020.

[29] D. F. Surianto and D. F. Surianto, “Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction,” *Brilliance: Research of Artificial Intelligence*, vol. 4, no. 2, pp. 964–972, Mar. 2025, doi: 10.47709/brilliance.v4i2.5547.

[30] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, “Improving Text Embeddings with Large Language Models,” 2024.