information

Article

Exploring Community Awareness of Mangrove Ecosystem Preservation through Sentence-BERT and *K*-Means Clustering

Retno Kusumaningrum [1,*], Selvi Fitria Khoerunnisa [2], Khadijah Khadijah [1] and Muhammad Syafrudin [3,*]

[1] Department of Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang 50275, Indonesia
[2] Master of Information System, School of Postgraduate Studies, Universitas Diponegoro, Semarang 50241, Indonesia; selvifitria@student.undip.ac.id
[3] Department of Artificial Intelligence and Data Science, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: retno@live.undip.ac.id (R.K.); udin@sejong.ac.kr (M.S.)

Abstract: The mangrove ecosystem is crucial for addressing climate change and supporting marine life. To preserve this ecosystem, understanding community awareness is essential. While latent Dirichlet allocation (LDA) is commonly used for this, it has drawbacks such as high resource requirements and an inability to capture semantic nuances. We propose a technique using Sentence-BERT and *K*-Means Clustering for topic identification, addressing these drawbacks. Analyzing mangrove-related Twitter data in Indonesian from 1 September 2021 to 31 August 2022 revealed nine topics. The visualized tweet frequency indicates a growing public awareness of the mangrove ecosystem, showcasing collaborative efforts between the government and society. Our method proves effective and can be extended to other domains.

Keywords: social media analysis; topic modeling; *K*-Means clustering; Sentence-BERT; mangrove awareness

Citation: Kusumaningrum, R.; Khoerunnisa, S.F.; Khadijah, K.; Syafrudin, M. Exploring Community Awareness of Mangrove Ecosystem Preservation through Sentence-BERT and *K*-Means Clustering. *Information* **2024**, *15*, 165. https://doi.org/10.3390/info15030165

Academic Editor: Leonardo Nascimento Ferreira

Received: 23 February 2024
Revised: 8 March 2024
Accepted: 12 March 2024
Published: 14 March 2024

1. Introduction

The use of microblogging in recent years has increased along with the emergence of various social media platforms. One of the microblogging platforms that users widely use is X (formerly known as Twitter). Microblogging allows users to utilize the space provided to create short messages or posts that are easy to understand so that the audience can receive concise information. The straightforward content of microblogging means that creating a microblog takes a short time, making it suitable for conveying time-sensitive information. In addition, from the reader's perspective, concise content from microblogging increases readership. Therefore, microblogging is the best choice for strengthening brand engagement or awareness.

Furthermore, what is meant by brands in the discussion of microblogging content does not only include trade or industrial commodity brands but also other things such as political figures, government policies, social conditions, the environment, etc. One environmental condition that needs to be known about in terms of the level of public awareness regarding the importance of its survival is the mangrove ecosystem. This is because mangroves are a vital habitat for (i) protecting coastal areas from abrasion, (ii) withstanding storms, (iii) filtering harsh pollutants, and (iv) providing living and spawning places for various types of marine biota. In addition, mangroves can provide various food sources for existing species.

Several techniques are widely implemented to understand public or community awareness using Twitter data, e.g., the following:

- Implementing sentiment analysis to find out how much the public has a negative opinion about a condition so that its importance is ignored or how much the public has a positive opinion about a condition so that the public becomes more aware of its importance [1–6];

- Examining changes in the use of hashtags, trending hashtags, and the frequency of related tweets [7,8];
- Implementing topic modeling [2,4–6];
- Matching the document with a lexicon of awareness expression markers [9];
- Extracting co-occurring or frequently occurring keywords from the document [4,10]

Based on the previous explanation, sentiment analysis is the most commonly applied technique. The implementation of sentiment analysis is limited to how people's opinions are linked to their level of awareness, without knowing what topics people have or have yet to understand. This drawback can be overcome by implementing topic modeling. The most widely implemented topic modeling algorithm is latent Dirichlet allocation (LDA). However, implementing LDA has several drawbacks, i.e., it requires many resources and a high computing time, and is unable to capture semantic nuances in a natural language. Therefore, we propose a simple technique for identifying abstract topics in a corpus by combining a robust word-embedding model, Sentence-BERT, and *K*-Means clustering. Sentence-BERT has several advantages, such as its ability to derive semantically meaningful sentence embeddings for specific tasks, including clustering and semantic similarity comparison [11].

Sentence-BERT represents a refinement of the BERT network, a type of large language model (LLM) employed for word embedding. Like other prominent LLM models such as Generative Pre-Trained Transformer (GPT) and Large Language Model Meta AI (LLaMA), BERT utilizes the transformer architecture. Notably, BERT distinguishes itself as a lightweight model from its peers. Moreover, BERT operates as an encoder-only transformer, ideal for tasks demanding a comprehensive comprehension of input text, such as trend topic analysis. In contrast, GPT and LLaMA function as decoder-only transformers, which excel in generative tasks. One of BERT's key strengths lies in its implementation of bidirectional training, enabling the capture of rich contextual information from both the left-to-right and right-to-left directions. Consequently, BERT demonstrates an enhanced ability to grasp the nuanced meaning and context of sentences compared with GPT, which operates solely in a unidirectional manner.

On the other hand, *K*-Means clustering has several advantages, such as being relatively simple to implement, robust, highly efficient, and applicable to various data types [12]. Hence, Sentence-BERT and *K*-Means clustering can be implemented as alternative techniques for identifying topics in a corpus. Moreover, by identifying the frequency of appearance of each topic each month and choosing the proper visualization technique, topic trend analysis regarding public awareness of the mangrove ecosystem can be obtained from Twitter data.

The rest of this paper is organized as follows. Section 2 describes the methodology employed, consisting of four main stages: Twitter scraping, text preprocessing, word embedding using Sentence-BERT, and trend topic analysis based on *K*-Means clustering. A detailed explanation of the results and their analysis are provided in Section 3. Finally, Section 4 concludes the results and suggests the direction of future work.

2. Related Works

2.1. Bidirectional Encoder Representation from Transformer (BERT)

BERT is a pre-trained language representation model developed by Devlin [13]. BERT is a transformer-based architecture in which a bidirectional encoder represents this architecture to train word representations based on context [14,15]. The BERT architecture was developed to train masked language modeling and sentence prediction embeddings.

Furthermore, the advantage of BERT is that there is a feature-based approach; thus, BERT can be scaled to large data sets. This approach is known as contextualized word embedding. Each word is mapped in vector space, and words with the same meaning are close to each other [16].

2.2. Sentence-BERT

BERT's feature-based approach maps each word into vector space. The same logic is applied in sentences; the simplest way is to take the average number of word vectors in a sentence. However, contextual word embeddings embed words in different contexts in supervised learning; they use weights with scalar mixing to obtain sentence embeddings. However, this is challenging in tasks involving semantic textual similarity (STS), such as clustering.

Therefore, Reimers and Gurevych [11] proposed Sentence-BERT to produce fixed-sized sentence embeddings for individual sentences by adding a pooling operation. Sentence-BERT is a modification of the BERT network, and is trained using a Siamese network and triplet network structure to obtain semantically meaningful sentence embeddings. Its architecture visualization can be seen in Figure 1.

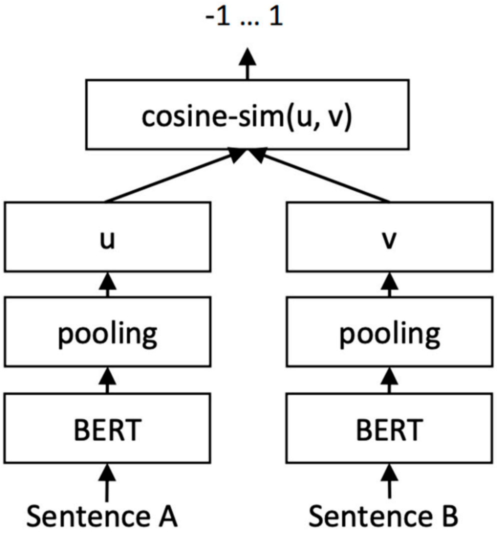

Figure 1. Sentence-BERT architecture.

Sentences are entered into the Transformers model for feature-based-approach tasks using pre-trained BERT. Then, the output vector is calculated based on the average of all vectors; the idea is to add a mean pooling operation that can represent sentences well but still maintain semantic similarity.

2.3. K-Means Clustering

K-Means is a clustering algorithm designed to partition unlabeled data into several clusters. *K*-Means requires an initial value of *K* to define the number of clusters. *K*-Means aims to separate data into groups with the same variation by minimizing an objective function as follows [14]:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \tag{1}$$

where r_{nk} is the membership value of data x_n in the k-th cluster, x_n is a data point, and μ_k is the centroid of the k-th cluster. The number of items in the dataset is N, and the number of clusters is K.

One challenging task in implementing *K*-Means Clustering is determining the number of clusters (K) since its value is generally unknown in actual applications. To resolve that task, cluster validity indices are commonly employed to estimate the number of clusters. Those indices are categorized into two major categories [17], namely, the following:

- External indices: It evaluates clustering results by comparing cluster memberships assigned by a clustering algorithm with previously known knowledge, such as class labels;

- Internal indices: It evaluates the goodness of a cluster structure by focusing on the intrinsic information on the data itself, such as its similarity.

Many internal indices, such as the elbow method, Dunn's index, the Davies–Bouldin index, silhouette width, the Calinski and Harabasz index, and gap statistics, are commonly implemented. The elbow method is an empirical method that is simple and easy to implement. This method plots the explained variations through the number of clusters, and it picks the elbow curve to obtain the optimal number of clusters based on the sum of squared errors within-cluster (SSEWC) of all data points to represent the quality of aggregation between data points in the same cluster and separation between clusters [18].

3. Methodology

This research consisted of four main stages: Twitter scrapping, text preprocessing, word embedding using Sentence-BERT, and trend topic analysis based on *K*-Means Clustering. The illustration of the research methodology can be seen in Figure 2.

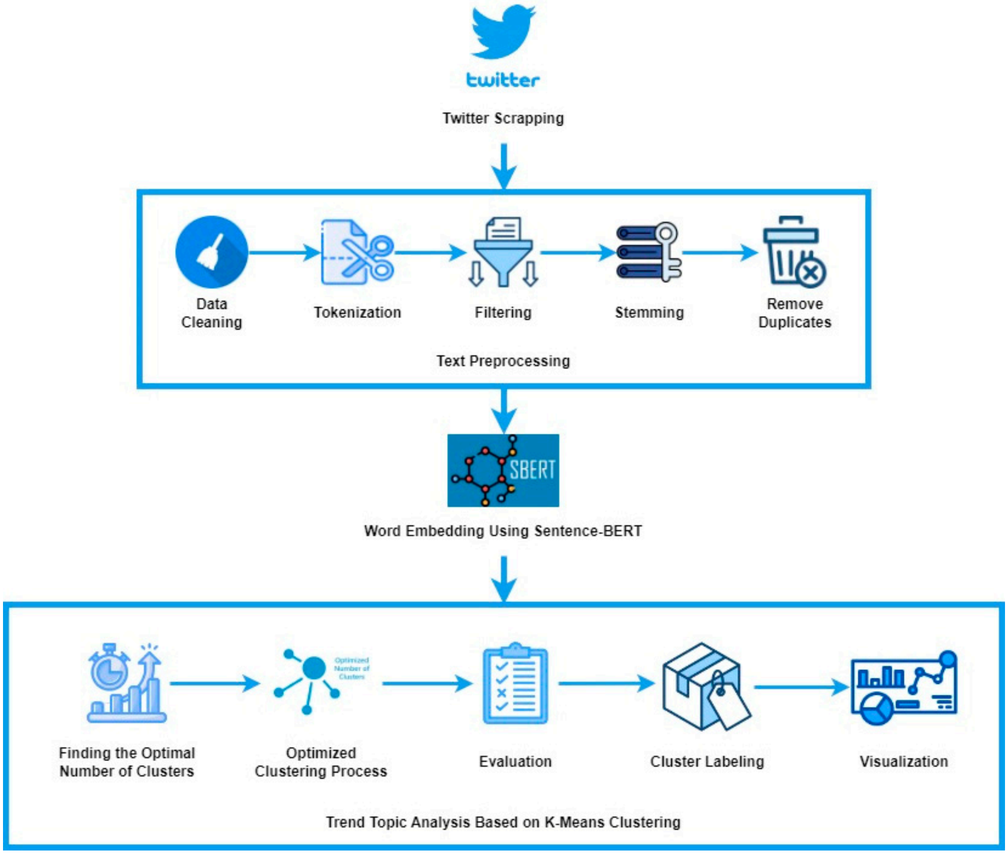
twitter
Twitter Scrapping
Data Cleaning
Tokenization
Filtering
Stemming
Remove Duplicates
Text Preprocessing
SBERT
Word Embedding Using Sentence-BERT
Optimized Number of Clusters
Finding the Optimal Number of Clusters
Optimized Clustering Process
Evaluation
Cluster Labeling
Visualization
Trend Topic Analysis Based on K-Means Clustering

Figure 2. Research methodology.

3.1. Twitter Scrapping

Twitter scrapping is a technique applied for dataset collection in this study. The employed data are specific Twitter data on mangrove and marine ecosystems written in Indonesian. Data were taken from 1 September 2021 to 31 August 2022, with about 27,371 tweets.

The tool employed for collecting the Twitter data is a snscrape libraries scraper. The advantage of the snscrape library is that there is no restriction on time for scraping, as enforced by Twitter API [19]. Since snscrape is a multilanguage library, we must set the "lang" parameter to "id". The scrapped data are subsequently stored in a data frame. Four labels of the data frame include the date, ID, username, and content. Since this study implements the *K*-Means Clustering algorithm, data frame labels have no class or output;

thus, it is an unsupervised learning algorithm. Table 1 shows a few tweet data examples from our dataset.

Table 1. Examples from dataset.

No.	Tweet in Indonesian (in English)
1	*Keberadaan hutan mangrove dapat mencegah terjadinya abrasi dan erosi pantai* (The existence of mangrove forests can prevent coastal abrasion and erosion)
2	*Pesona kawasan taman hutan raya mangrove jadi venue penting di even G20 Bali* (The beauty of the mangrove forest park area is essential to the G20 Bali event)
3	*Ayo Sob, kita dukung rehabilitasi mangrove yang melibatkan masyarakat di Provinsi Kepulauan Bangka Belitung* (Come on, friend, let us support mangrove rehabilitation involving the Bangka Belitung Islands Province community)
4	*Kelompok Rehabilitasi Ekosistem Pesisir Kampung Yensawai Ajak Wisatawan Asing Tanam Karang, Lamun, Terumbu Karang dan Mangrove* (Yensawai Village Coastal Ecosystem Rehabilitation Group Invites Foreign Tourists to Plant Coral, Seagrass, Coral Reefs and Mangroves)
5	*Mangrove adalah salah satu pembahasan penting dalam Presidensi G20 Indonesia 2022, mencakup issu lingkungan dunia* (Mangroves are one of the critical discussions in Indonesia's 2022 G20 Presidency, covering world environmental issues)
6	*Selamat Hari Mangrove Internasional, Ayo Jaga hutan Mangrove kita* (Happy International Mangrove Day! Let's protect our mangrove forests)
7	*Restorasi mangrove teguhkan komitmen hadapi dampak perubahan iklim* (Mangrove restoration strengthens commitment to facing the impacts of climate change)

3.2. Text Preprocessing

Text preprocessing is the most significant procedure in computational linguistics. This stage's goal is knowledge mining, which includes transforming raw data and making them ready to be entered into a machine learning model [20]. This stage consists of five steps, depicted in Figure 1, including data cleaning, tokenization, filtering, stemming, and removing duplicate data.

3.2.1. Data Cleaning

As mentioned before, Twitter is the source of the collected dataset in this study. It is widely known that social media data, specifically from Twitter, have much noise, so data cleaning was required to remove unnecessary elements such as punctuation and symbols. This step is divided into several sub-processes, namely the following:

- Remove the hashtag by replacing it with "<hashtag>,";
- Remove the link or web address by replacing it with "<link>,";
- Remove the handler or mention by replacing it with "<user>.";
- Remove the emoticon or emoji by replacing it with "<emoji>.";
- Remove single characters, numbers, punctuation, and special characters that are unnecessary in analysis.

3.2.2. Tokenization

Tokenization is the procedure for segmenting sentences into words, the results of which are called tokens [21]. Sentences are truncated based on whitespaces such as tabs and blanks, and punctuation symbols such as commas, semicolons, periods, and colons [22]. This stage eliminates unnecessary punctuation, spaces, and characters.

3.2.3. Filtering

Filtering is a procedure for selecting important words resulting from tokenization. An essential task in filtering is the algorithm removing important words (stop list) and saving important words (word list) resulting from tokenization. The type of filtering used in this study was stopword removal. Stopword removal removes words that do not contribute to the main text features, thereby reducing dimensionality [23].

3.2.4. Stemming

Stemming is the stage for changing all words into base words. The stemming process removes all word affixes (affixes) in derivative words. This process uses the literary library, a stemmer algorithm specifically for Indonesians. This process needs to be paid attention to because of the differences in the syntax of each language.

3.2.5. Remove Duplicates Data

After the four previous steps, some duplicate data were obtained. Subsequently, these duplicate data were removed to optimize the performance of the following stages. The number of datasets ready for use after eliminating duplicates was 26,671 tweets. Table 2 shows the detailed statistics of the final dataset, where the datasets are grouped per month using the month ID to distinguish the time range. It is important to emphasize that the quantity of Tweets gathered is not within our purview, as they are autonomously generated by Twitter users, uninfluenced by our study. Notably, the temporal dimension (month and year) of data collection was not incorporated into our analysis. Therefore, fluctuations in monthly and yearly tweet volumes are unlikely to exert an influence on our findings.

Table 2. Dataset distribution for each month.

Month_ID	Period	Number of Tweets
1	September 2021	1467
2	October 2021	1390
3	November 2021	1370
4	December 2021	922
5	January 2022	1417
6	February 2022	2762
7	March 2022	2885
8	April 2022	2670
9	May 2022	2879
10	June 2022	4307
11	July 2022	2012
12	August 2022	2590
	Total number of Tweets = 26,671	

3.3. Word Embedding Using Sentence-BERT

The word embedding technique employed in this sentence is Sentence-BERT, a modification of BERT developed for unsupervised tasks such as clustering. Two types of pre-trained BERT are used as a basis for word embedding. The first is distiluse-base-multilingual-cased-v2 [24]. This pre-trained multilingual model has been trained in over 50 languages (including Indonesian), with 6 layers, 768 dimensions, 128 max sequences, 12 heads, and 66M as parameters.

The second pre-trained model is IndoBERTweet [25], a retrained model for Indonesian Twitter, which was trained on the monolingually trained IndoBERT model. This model was trained with 12 hidden layers (dimensional = 768), 12 attention heads, three feed-forward hidden layers (dimensional = 3072), and 409 M word tokens.

The dataset resulting from preprocessing is represented in vector form using the Sentence-BERT method and the two pre-trained BERT methods mentioned. The word embedding stages include three types: token embedding, segment embedding, and position embedding [26].

3.4. Trend Topic Analysis Based on K-Means Clustering

As depicted in Figure 1, there are five steps in the trend topic analysis stage: finding the optimal number of clusters, optimizing the clustering process, evaluation, cluster labeling, and trend topic visualization. The following sub-sections explain the details of each step.

3.4.1. Finding the Optimal Number of Clusters

Most clustering methods rely on a predetermined number of clusters, although in actual circumstances, the number of clusters cannot be predicted in advance [27]. As a starting point, to determine the optimal number of clusters, we initialized the number of clusters from 2 to 50, having yet to gain prior knowledge about the optimal number of clusters in the dataset used. Each cluster was fitted to the *K*-Means Clustering algorithm, which was then evaluated using the elbow method.

The elbow method is the most common method used to determine the number of clusters. According to Shi et al. [27], the initial idea of this elbow method is to determine that the number of clusters (*K*) is equal to 2 as the initial number of clusters and then increase it to the specified maximum *K* value. The optimal *K* value corresponding to the plateau will later be determined. This method captures variations in the dataset using a within-cluster sum of squared errors (WSS). The variation value decreases sharply before the number of clusters exceeds the optimal *K* value. The optimal point is determined from the corner point of the sharp change in variation in the cluster.

3.4.2. Optimizing the Clustering Process

The optimal number of clusters obtained from the elbow method is used as a parameter in running the optimal clustering model. This number of clusters is used to initialize *K*-Means Clustering. The labels resulting from fitting the *K*-Means model are used as topic clusters for each document, which are later analyzed to identify relevant subjects.

3.4.3. Evaluation

The optimal model that has been run is then evaluated for accuracy. The evaluation metrics used are the silhouette score and coherence score. The silhouette score uses the average distance between one data point and other data points in the same cluster, and the average distance between different clusters and other data points. In contrast, the coherence score measures the score of a topic based on the semantic similarity between words that have a score high in the topic.

3.4.4. Cluster Labeling

The cluster labeling process was performed by generating a WordCloud for each cluster. Each WordCloud consists of the ten most frequently appearing words in the respected cluster. The manual interpretation was conducted to infer the cluster label based on WordCloud.

3.4.5. Trend Topic Visualization

After all the tweets have known topics, the next step is to visualize the trends in these topics over time. The number of tweets appearing each month for each topic are calculated, and then a visualization is created using a stacked bar chart. This chart type is selected based on several conditions (i) how discussions regarding the mangrove ecosystem fluctuate occasionally on Twitter; (ii) fluctuations in the discussion for each topic related to the mangrove ecosystem from time to time, and a comparison of the volume of each topic per period.

4. Results and Discussions

This section describes the results and discussion of this research. As previously explained, this research used 26,671 tweets about mangrove and marine ecosystems from Twitter, word embedding with Sentence-BERT, and *K*-Means Clustering. The detailed results of this study will be explained in the following sub-sections.

4.1. Optimal Number of Clusters

This study determined the optimal number of clusters based on the elbow method. By initializing the number of clusters from 2 to 50, the elbow method was used to indicate the cost function value produced by different K values in the two pre-trained BERT models. Figure 3 shows the elbow method plot on pre-trained distiluse-base-multilingual-cased-v2. Based on Figure 3, the optimal number of clusters is nine since there is a sharp decrease for $K < 9$, and for the inertia value, steady decreases occur for $K > 9$. Meanwhile, the optimal value in the pre-trained IndoBERTweet, as depicted in Figure 4, shows that the optimal number of clusters is seven. For the same reason, there is a sharp decline in $K < 7$; then, a break occurs, and the graph starts to slope at $K > 7$.

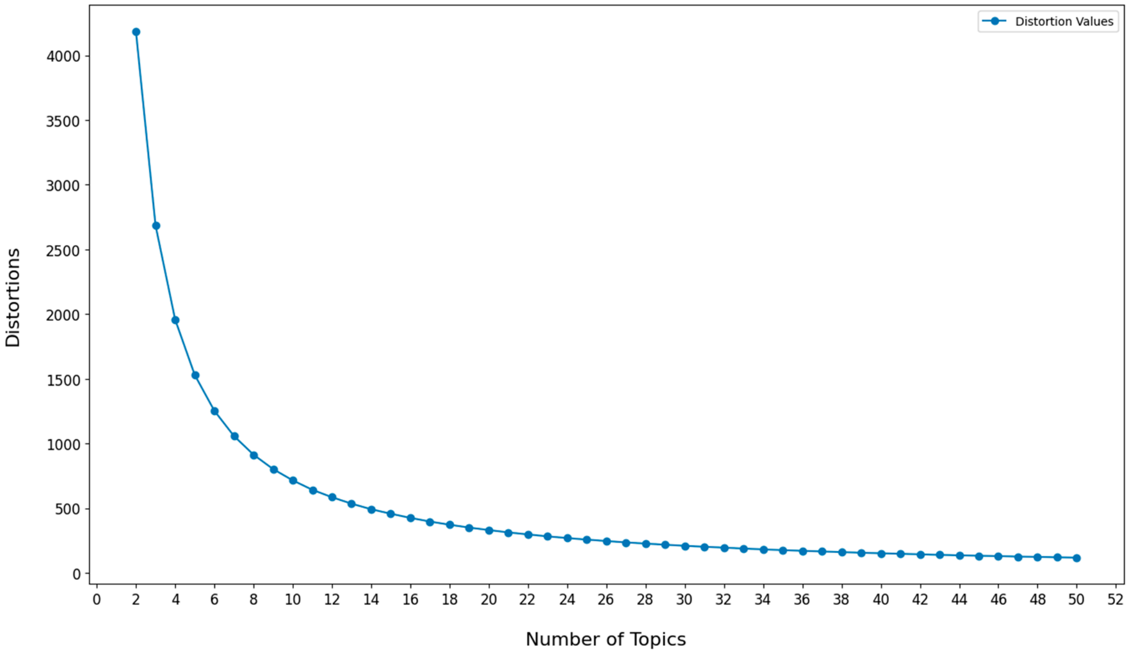

Figure 3. Elbow for pre-trained distiluse-base-multilingual-cased-v2.

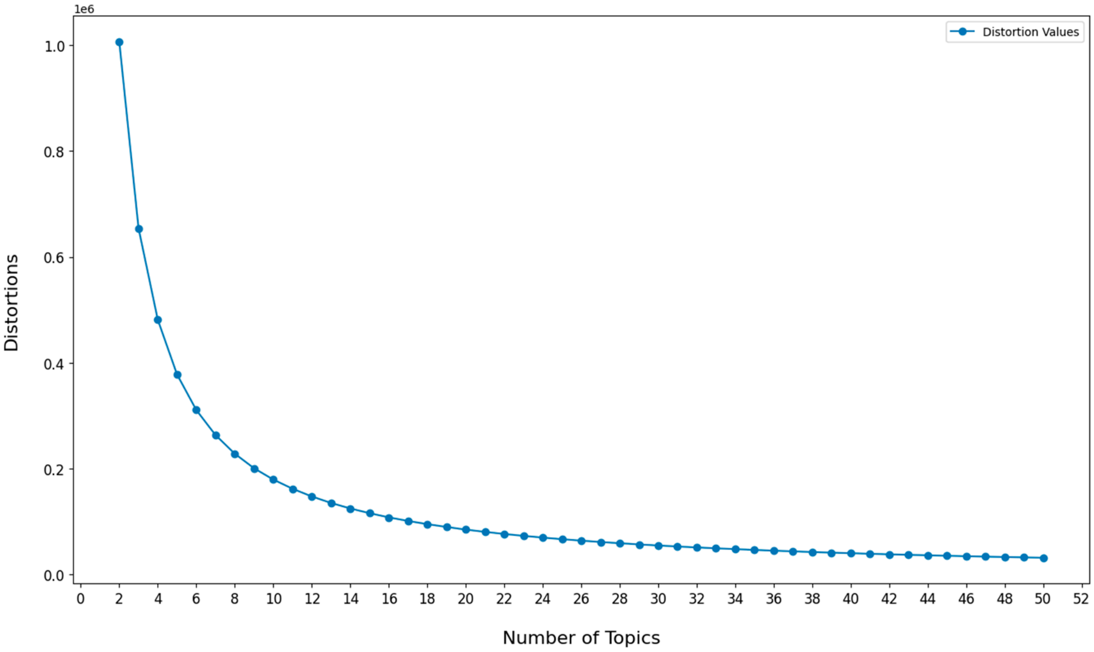

Figure 4. Elbow for pre-trained IndoBERTweet.

4.2. K-Means Clustering Based on Optimal Number of Clusters

After the best number of clusters is known, the next step is to rerun K-Means Clustering with K values of 9 and 7 for distiluse-base-multilingual-cased-v2 and IndoBERTweet, respectively. Subsequently, the clustering results are evaluated using the coherence and silhouette scores, as explained in Table 3.

Table 3. Evaluation results of *K*-Means Clustering.

Pre-Trained Model	Coherence Score	Silhouette Score
distiluse-base-multilingual-cased-v2	0.453	0.036
IndoBERTweet	0.383	0.018

It can be seen that the silhouette model values for each pre-trained are 0.036 and 0.018. The silhouette value ranges from −1 to 1, which means that the silhouette score results of the two pre-trained above are close to zero. Silhouette values that are close to zero indicate that the data in a cluster have a relatively low level of similarity compared with data in other clusters; the cluster boundaries may not be evident, so some data have quite significant similarities with data in other clusters. We analyzed this, starting from Twitter, which contains complex and unstructured data. Data can contain many language and communication styles, so this variability makes it challenging to group tweets into clear clusters. In addition, one tweet can have several topics, which can be captured by implementing hard clustering, such as *K*-Means Clustering.

In simple terms, the coherence score measures the degree of semantic similarity between data in one cluster. It can also be interpreted as an approach to measuring the model in reflecting the characteristics of each cluster. Furthermore, the coherence scores obtained in the two models are 0.453 and 0.383. The coherence score value ranges from 0 to 1, which is quite good and indicates that the words in the cluster have an excellent level of semantic similarity due to the implementation of Sentence-BERT as a word-embedding model.

Furthermore, it can be seen that the results of the multilingual pre-trained model from distiluse-base-multilingual-cased-v2 have better evaluation results than do those from IndoBERTweet based on both coherence and silhouette scores. IndoBERTweet is a monolingual BERT pre-trained model explicitly trained for the Indonesian Twitter domain. This model should have better evaluation results. However, its results are different from the results obtained in this study. This result is influenced by the condition in which IndoBERTweet was trained with the economy, health, education, and government data over one year. From this statement, the type of topic data used in choosing a pre-trained BERT model is essential. Although the monolingual model may have a better understanding of the local language context, the representation of the use of words that are pre-trained is also influenced by the subject or topic of the data contained therein. Moreover, multilingual models have larger model sizes because they are trained in many languages to handle the linguistic variations in Twitter data. Another possibility is that multilingual models are better at generalizing across languages, where there may be uncommon words that have similarities to those in other languages.

4.3. Cluster Labeling

As explained before, the best result is the cluster produced based on the distiluse-base-multilingual-cased-v2 pre-trained model. After the results of the clustering evaluation are obtained, cluster labeling is manually performed by generating a WordCloud. The WordCloud consists of ten words with the highest frequency of occurrence, as depicted in Figure 5. Subsequently, manual interpretation is carried out to conclude the cluster labeling. The cluster labeling results can be seen in Table 4.

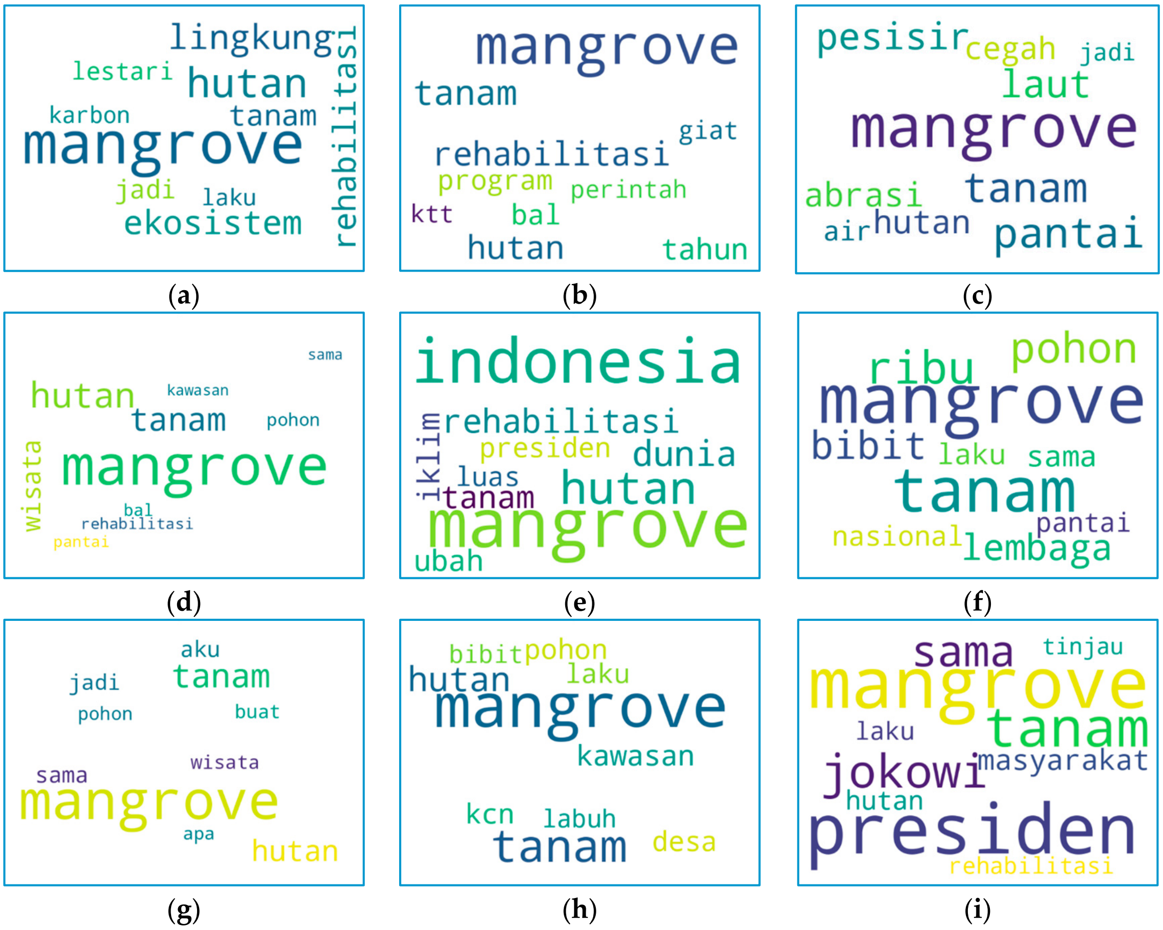

Figure 5. WordCloud visualization for topic labeling for pre-trained IndoBERTweet. (**a**). Topic 1, (**b**). Topic 2, (**c**). Topic 3, (**d**). Topic 4, (**e**). Topic 5, (**f**). Topic 6, (**g**). Topic 7, (**h**). Topic 8, (**i**). Topic 9.

Table 4. Label and keyword for each cluster.

Topic ID	Label	Keywords
1	Benefits of mangrove ecosystem rehabilitation	*mangrove, lingkungan, rehabilitasi, lestari, hutan, karbon, tanam, jadi, ekosistem* mangrove, environment, rehabilitation, sustainable, forest, carbon, planting, ecosystem
2	Government program for mangrove forests in the context of the G-20 Summit	*mangrove, tanam, rehabilitasi, giat, tanam, program, pemerintah, ktt, hutan, tahun* mangrove, planting, rehabilitation, active, planting, program, government, summit, forest, year
3	The function of mangrove forests is to prevent abrasion	*mangrove, pesisir, cegah, jadi, laut, abrasi, tanam, air, hutan, pantai* [mangrove, coastal, prevent, so, sea, abrasion, planting, water, forest, beach]
4	Mangrove forest ecotourism	[*mangrove, hutan, Kawasan, pohon, sama, tanam, wisata, rehabilitasi, pantai*] mangrove, forest, area, tree, same, planting, tourism, rehabilitation, beach
5	Mangrove forests and their relation to climate change	*mangrove, indonesia, iklim, presiden, luas, dunia, hutan, tanam, ubah, luas* mangrove, indonesia, climate, president, area, world, forest, plant, change, area
6	Public center of mangrove ecosystem	*mangrove, pohon, ribu, bibit, laku, sama, tanam, Pantai, nasional, lembaga* mangrove, tree, thousand, seeds, sell, same, plant, beach, national, institution
7	Replanting mangrove forests	*mangrove, tanam, hutan, wisata, hutan, pohon, apa, sama, jadi, buat* mangrove, planting, forest, tourism, forest, tree, what, same, so, make
8	Mangrove tourist village	*mangrove, bibit, hutan, pohon, laku, kawasan, desa, labuh, tanam, kcn* mangrove, seedling, forest, tree, practice, area, village, anchor, planting, kcn
9	President's visit to mangrove forests in the context of the G-20 Summit	*mangrove, tinjau, laku, tanam, masyarakat, jokowi, presiden, hutan, rehabilitasi, sama* mangrove, review, sell, plant, community, jokowi, president, forest, rehabilitation, same

Subsequently, the subjects in each cluster label can be explained as follows:

- Topic 1 contains tweets about mangrove rehabilitation, an ecosystem improvement that benefits society and the environment (e.g., by reducing carbon emissions);
- Topic 2 contains tweets about the government's program for rehabilitating mangrove forests in preparation for the G-20 Summit;
- Topic 3 contains tweets about real community action in preventing abrasion by planting mangrove trees along the coast;
- Topic 4 contains tweets about the potential for coastal tourism areas and mangrove forests;
- Topic 5 contains tweets about actions to build mangrove centers in Indonesia to deal with climate change;
- Topic 6 contains tweets about the actual actions of the national government in planting mangrove tree seedlings;
- Topic 7 contains tweets about actual community action in replanting mangrove forests;
- Topic 8 contains tweets about the potential of mangrove tourism villages;
- Topic 9 contains tweets about the President's visit to the mangrove forest in preparation for the G-20 Summit.

Based on Figure 5 and the explanation of the subject above, it can be seen that several topics are related. As with topic 1, "Benefits of mangrove ecosystem rehabilitation," and topic 7, "Replanting mangrove forests," the main aim is to rehabilitate and replant mangrove forests. Still related to this, topics 3 and 5 discuss the benefits of mangrove forests, such as preventing abrasion and environmental conservation efforts. Mangrove roots can prevent coastal erosion, and mangrove forests, the most significant contributors to oxygen and absorb carbon emissions, can prevent world climate change, especially those in Indonesia. Moreover, the government also supports strategic efforts to preserve mangrove forests. This can be seen from topics 2 and 9. These two topics discuss cooperation programs between countries at the G-20 Summit Meeting. Table 5 shows an example of tweets corresponding to their topic IDs.

Table 5. Example of tweets based on topic.

Example of Tweet in Indonesian (in English)	Topic ID
Rehabilitasi mangrove, konservasi maupun tata kelolanya sangat penting sebagai suistainable development (Mangrove rehabilitation, conservation, and management are crucial for sustainable development)	1
Makanya bakal ada showcase konservasi di Bali loh selama rangkaian acara #KickOffG20, showcase ini jg jadi bagian rehabilitasi &amp; konservasi mangrove (That's why there will be a conservation showcase in Bali during the #KickOffG20 event series; this showcase will also be part of mangrove rehabilitation and conservation.)	2
Tanam Mangrove, Kurangi Potensi Abrasi di Daerah Pesisir Pantai (Planting mangroves reduces the potential for abrasion in coastal areas.)	3
Sejauh mata memandang, hamparan mangrove yang tampak menghijau menjadi penyejuk mata di siang hari yang cukup terik di kawasan ekowisata Karongsong di pesisir pantai Indramayu, Jabar. (The verdant mangroves extend as far as the eye can see, refreshing the eyes on a hot day in the Karongsong ecotourism area on the coast of Indramayu, West Java).	4
Mangrove yang kaya karbon di kepulauan Indonesia harus menjadi komponen strategi prioritas tinggi untuk mitigasi perubahan iklim (Carbon-rich mangroves in the Indonesian archipelago should be a high-priority strategy component for climate change mitigation)	5
Pemerintah sudah merehabilitasi mangrove di lahan 110 ribu hektar dari rencana yang sudah ditargetkan 600 ribu hektar (The government has rehabilitated mangroves on 110 thousand hectares of land from the planned target of 600 thousand hectares)	6
Ini merupakan penanaman #mangrove yang dilakukan minggu lalu, dalam rangka menyambut peringatan Hari Mangrove Sedunia di pesisir Pantai Tirang, Jawa Tengah (This is a #mangrove planting that was carried out last week to welcome the commemoration of World Mangrove Day on the coast of Tirang Beach, Central Java.)	7
The Sungai Kupah Tourism Village management is holding a Mangrove Edu camp throughout West Kalimantan to celebrate World Mangrove Day. #mangroveeducamp (The Sungai Kupah Tourism Village management is holding a Mangrove Edu camp throughout West Kalimantan to celebrate World Mangrove Day. #mangroveeducamp)	8
Jelang G20 Presiden Tinjau Hutan Mangrove (Ahead of G20, President Visits Mangrove Forests)	9

4.4. Trend Topic Visualization

As explained, this study implements a stacked bar chart to visualize the trend topic. The chart can be seen in Figure 6. There is an increase in the number of tweets in October 2021, as depicted in Figure 6. This is in line with the increase in tweet volume on topic nine about the President's visit to the mangrove forest in preparation for the G-20 Summit. Even though the G-20 Summit was held on 15–16 November 2022, with the government visiting the mangrove forest and mentioning that the mangrove forest will be one of the venues, the community started to show interest. The government's mangrove rehabilitation program will continue to be carried out beyond the G-20 event. Many national governments and the private sector participate in planting mangrove tree seedlings, as shown in topic 6.

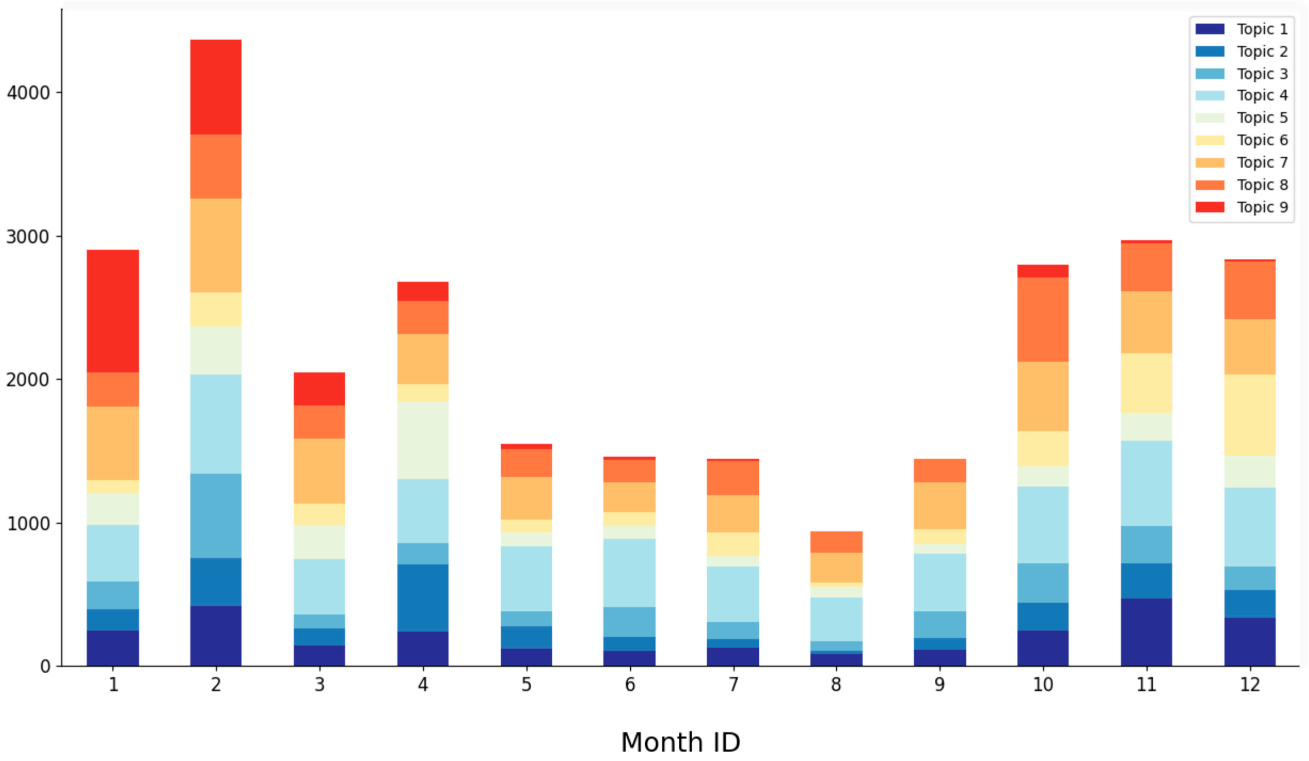

Figure 6. Trend topic distribution.

Apart from the efforts of the community and government in rehabilitating mangrove forests, the community also uses Twitter to promote mangrove forest tourism in their respective areas. This is the main topic because it has the most significant percentage of data. In line with this, topic eight also explains the potential of mangrove tourism villages.

Based on the statements above, the government and society continue to strive to create a synergy to continue to preserve mangrove forests. We observed that both are trying to increase awareness of the importance of mangrove forests by continuously trying to demonstrate the mangrove care movement. This can be seen from the distribution of topics related to mangrove rehabilitation (topics 3 and 5), which continuously appear every month. Even though most of the mangrove rehabilitation and replanting programs are initiated by the government or agencies, the community also actively participates. Therefore, consistent invitation efforts are needed so that this effort can have long-term benefits.

5. Conclusions

This study proposed the combination of Sentence-BERT and *K*-Means Clustering to identify topics in a corpus. Subsequently, trend topic analysis was performed via visualization using a stacked bar chart that was used to inform the tweet frequency for each topic every month. The trend topic analysis was conducted on mangrove-related

Twitter tweets in the Indonesian language for a certain period from 1 September 2021 to 31 August 2022.

There were nine topics obtained: benefits of mangrove ecosystem rehabilitation, the government program for mangrove forests in the context of the G-20 Summit, the function of mangrove forests is to prevent abrasion, mangrove forest ecotourism, mangrove forests and their relation to climate change, the public center of the mangrove ecosystem, replanting mangrove forests, a mangrove tourist village, and the President's visit to mangrove forests in the context of the G-20 Summit. Based on the stacked bar chart visualization, it can be seen that the government and society continue to create a synergy in preserving mangrove forests by continuously trying to advocate for the mangrove care movement so that public awareness of the importance of the mangrove ecosystem continues to grow.

In future research, employing the combined approach of Sentence-BERT and LDA could serve as a viable alternative for identifying topics within a corpus. Utilizing Sentence-BERT is anticipated to address the limitations of LDA by enabling the capture of semantic nuances inherent in natural language. Concurrently, the implementation of LDA is expected to yield more adaptable topics, thereby accommodating the emergence of diverse themes in tweets based on real-world data conditions. Additionally, recent advancements in natural language processing (NLP) models, such as LLaMa and GPT, merit investigation to enhance the effectiveness of topic discovery and analysis. Furthermore, the potential for expanding data collection from alternative microblogging platforms warrants consideration for future exploration.

Author Contributions: Conceptualization, R.K. and M.S.; methodology, R.K. and S.F.K.; software, S.F.K. and K.K.; validation, S.F.K. and K.K.; formal analysis, R.K. and M.S.; data curation, S.F.K. and K.K.; writing—original draft preparation, R.K., S.F.K., and K.K.; writing—review and editing, M.S.; visualization, S.F.K. and K.K.; supervision, R.K. and M.S.; funding acquisition, R.K. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Universitas Diponegoro under Grant of Research for International Publication, Number 753-26/UN7.D2/PP/V/2023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Reyes-Menendez, A.; Saura, J.R.; Alvarez-Alonso, C. Understanding #worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach. *Int. J. Environ. Res. Public. Health* **2018**, *15*, 2537. [CrossRef] [PubMed]
2. Karami, A.; Shah, V.; Vaezi, R.; Bansal, A. Twitter Speaks: A Case of National Disaster Situational Awareness. *J. Inf. Sci.* **2019**, *46*, 313–324. [CrossRef]
3. D'andrea, E.; Ducange, P.; Bechini, A.; Renda, A.; Marcelloni, F. Monitoring the Public Opinion about the Vaccination Topic from Tweets Analysis. *Expert Syst. Appl.* **2019**, *116*, 209–226. [CrossRef]
4. Boon-Itt, S.; Skunkan, Y. Public perception of the COVID-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill.* **2020**, *6*, e21978. [CrossRef] [PubMed]
5. Chandrasekaran, R.; Mehta, V.; Valkunde, T.; Moustakas, E. Topics, Trends, and Sentiments of Tweets about the COVID-19 Pandemic: Temporal Infoveillance Study. *J. Med. Internet Res.* **2020**, *22*, e22624. [CrossRef] [PubMed]
6. Bian, J.; Zhao, Y.; Salloum, R.G.; Guo, Y.; Wang, M.; Prosperi, M.; Zhang, H.; Du, X.; Ramirez-Diaz, L.J.; He, Z.; et al. Using social media data to understand the impact of promotional information on laypeople's discussions:a case study of lynch syndrome. *J. Med. Internet Res.* **2017**, *19*, e414. [CrossRef] [PubMed]
7. Patel, K.D.; Zainab, K.; Heppner, A.; Srivastava, G.; Mago, V. Using Twitter for diabetes community analysis. *Netw. Model. Anal. Health Inform. Bioinform.* **2020**, *9*, 36. [CrossRef]
8. Abbar, S.; Zanouda, T.; Berti-Equille, L.; Borge-Holthoefer, J. Using Twitter to Understand Public Interest in Climate Change: The Case of Qatar. In Proceeding of the Tenth International AAAI Conference on Web and Social Media Social Web for Environmental and Ecological Monitoring, Cologne, Germany, 17–20 May 2016; pp. 168–177.

9. Lenoir, P.; Moulahi, B.; Azé, J.; Bringay, S.; Mercier, G.; Carbonnel, F. Raising awareness about cervical cancer using twitter: Content analysis of the 2015 #smearforsmear campaign. *J. Med. Internet Res.* **2017**, *19*, e344. [CrossRef] [PubMed]
10. Xie, Q.; Zhou, Y.; Xin, L.; Qianqian, X.; Lucheng, H. Twitter Data Mining for the Social Awareness of Emerging Technologies. In Proceedings of the 2017 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 9–13 July 2017; pp. 1–10. [CrossRef]
11. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]
12. Wu, J. Cluster Analysis and *K*-Means Clustering: An Introduction. In *Advances in K-Means Clustering, A Data Mining Thinking*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 1–16. [CrossRef]
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
14. Subakti, A.; Murfi, H.; Hariadi, N. The performance of BERT as data representation of text clustering. *J. Big Data* **2022**, *9*, 15. [CrossRef] [PubMed]
15. Hu, W.; Xu, D.; Niu, Z. Improved *K*-Means Text Clustering Algorithm Based on BERT and Density Peak. In Proceedings of the 2021 2nd Information Communication Technologies Conference, ICTC 2021, Nanjing, China, 7–9 May 2021; pp. 260–264. [CrossRef]
16. Kaliyar, R.K. A Multi-layer Bidirectional Transformer Encoder for Pre-trained Word Embedding: A Survey of BERT. In Proceeding of the 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 29–31 January 2020; pp. 336–340.
17. Sinaga, K.P.; Yang, M.S. Unsupervised *K*-Means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
18. Sammouda, R.; El-Zaart, A. An Optimized Approach for Prostate Image Segmentation Using *K*-Means Clustering Algorithm with Elbow Method. *Comput. Intell. Neurosci.* **2021**, *2021*, 4553832. [CrossRef] [PubMed]
19. Bhatt, B. Scrape Twitter Data or Tweets in Python Using Snscrape Module. Available online: https://github.com/bhattbhavesh91/twitter-scrapper-snscrape (accessed on 3 September 2022).
20. George, L.; Sumathy, P. An integrated clustering and BERT framework for improved topic modeling. *Int. J. Inf. Technol.* **2023**, *15*, 2187–2195. [CrossRef] [PubMed]
21. Jeremy, N.H.; Suhartono, D. Automatic personality prediction from Indonesian user on twitter using word embedding and neural networks. *Procedia Comput. Sci.* **2021**, *179*, 416–422. [CrossRef]
22. Khan, A.; Shah, Q.; Uddin, M.I.; Ullah, F.; Alharbi, A.; Alyami, H.; Gul, M.A. Sentence embedding based semantic clustering approach for discussion thread summarization. *Complexity* **2020**, *2020*, 4750871. [CrossRef]
23. Zhu, L.; Luo, D. A Novel Efficient and Effective Preprocessing Algorithm for Text Classification. *J. Comput. Commun.* **2023**, *11*, 1–14. [CrossRef]
24. Reimers, N.; Gurevych, I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 4512–4525. Available online: https://github.com/facebookresearch/ (accessed on 15 December 2023).
25. Koto, F.; Lau, J.H.; Baldwin, T. INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 10660–10668. Available online: https://huggingface.co/huseinzol05/ (accessed on 15 December 2023).
26. Xie, Q.; Zhang, X.; Ding, Y.; Song, M. Monolingual and multilingual topic analysis using LDA and BERT embeddings. *J. Informetr.* **2020**, *14*, 101055. [CrossRef]
27. Shi, C.; Wei, B.; Wei, S.; Wang, W.; Liu, H.; Liu, J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J. Wirel. Commun. Netw.* **2021**, *2021*, 31. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.