Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.9 No.2 July 2024, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles

Dini Aryani[1], Ivana Lucia Kharisma[2], Alun Sujjada[3], Kamdan[4]

[1,2,3,4]*Informatics Department, Nusa Putra University, Sukabumi, Indonesia*

[2]ivana.lucia@nusaputra.ac.id (*)

[1,3,4][dini.aryani_ti20, alun.sujjada, kamdan]@nusaputra.ac.id

Received: 2024-06-12; Accepted: 2024-07-27; Published: 2024-08-03

Abstract— In 2024, Indonesia will hold simultaneous general elections dominated by the participation of young people, especially Generation Z and Millennials, who seek political information primarily through the Internet, highlighting the crucial role of digital media in shaping public opinion. Detik.com is actively reporting on the 2024 elections, evidenced by a special election subchannel. However, the lack of topic categorization in this subchannel makes it difficult for readers to find in-depth information, and tracking and analyzing the large volume of news articles published daily is a significant challenge. This study employs Topic Modelling techniques, specifically the BERTopic method, to analyze topics related to the 2024 elections from Kompas.com news articles. The dataset, sourced from the detik.com election sub-channel, was collected via scraping from September 1, 2023, to February 14, 2024, totalling 15,019 articles. The text preprocessing involves case folding, cleaning, tokenizing, and stopword removal. Topic modelling using BERTopic includes embeddings with sentence-transformers "distiluse-base-multilingual-cased-v1," dimensionality reduction with UMAP, clustering with K-Means using optimal $k=5$ value evaluated by Elbow, tokenizer with CountVectorizer, and weighting scheme using c-TF-IDF. Based on the Silhouette Score of 0.566 and Silhouette plot results, the clustering results using the K-Means model with a value of k equal to 5 produce good clustering with clear inter-cluster distances. For other evaluations, the SSE value of 70223.257 provides an overview of the cluster distribution, the Davies-Bouldin Index of 0.758 shows that the cluster has a relatively good level of inter-cluster separation with good closeness within the cluster, the Calinski-Harabasz Index of 20083.489 shows good and compact inter-cluster separation, and the Dunn Index of 0.003 shows outliers that cause overlapping clusters and lack of clear separation. The evaluation results show that implementing the K-Means model with a value of k equal to 5 again emphasizes that the clustering results are good. The modelling results show an average topic coherence value of 0.0902 and produce five main topics in the 2024 election news on Detik.com topic 0: about presidential and vice presidential candidates (5,215 articles) with the representation of the words *'ganjar'*, *'prabowo'*, *'anies'* and *'imin'*, topic 1: about general elections and related surveys (3,191 articles) with the representation of the words '2024', *'pemilu'*, *'pilpres'* and *'suara'*, topic 2: news about Joko Widodo President (2,604 articles), topic 3: news about presidential and vice presidential debates (2046 articles) with representations of the words *'presiden'*, *'jokowi'*, *'demokrat'* and *'politik'* and topic 4: news about the figure Gibran Rakabuming Raka and related issues (1963 articles) with representations of the words *'raka'*, *'rakabuming'*, *'nomor'* and *'urut'*. Using the results of this research, readers can gain insights into the most discussed issues and the attention given to key figures in the 2024 election news on the detik.com news portal.

Keywords— Topic modelling; BERTopic Method; Sentimen Analysis; 2024 Election; detik.com.

I. INTRODUCTION

On February 14, 2024, Indonesia held simultaneous general elections to determine leaders and legislative members at the district/city, provincial, and national levels, as well as representatives in the DPD-RI. Generation Z, born in 1995 to 2000, has voting rights of around 22.8%. Meanwhile, millennials, born between 1980 and 1994, have about 33.6% of the total vote [1], [2]. The 2024 election will be characterized by the dominant participation of young voters, namely Generation Z and Millennials. Both generations have grown up in the digital age, where access to the Internet and social media is an important part of their daily lives [3]. Generation Z and Millennials actively discuss, share news articles, and seek opinions from various online sources [4]. Thus, the Internet is the main source for millennials and Z generations to form political views and make decisions in elections. This interaction is then facilitated by online mass media presenting actual news that all groups can access. The availability and ease of accessing mass media online make this media a very popular choice for the public [5].

Detik.com is an online news site in Indonesia founded on July 9, 1998, by Budiono Darsono, Yayan Sopyan, Abdul Rahman, and Didi Nugrahadi. Since August 3, 2011, detik.com has been part of PT Trans Corporation, a subsidiary of CT Corp [6]. According to data from SimilarWeb in January 2024, Detik.com was ranked in the top 3 leading news and media publishers. Detik.com is actively reporting on the 2024 elections, as evidenced by a special sub-channel reporting on the elections. However, Detik.com does not categorize news based on specific topics in the Detik election subchannel. This makes it difficult for readers to find more in-depth information on the sub-channel. However, tracking and analyzing topics that appear in news coverage is challenging, especially given the large volume of news articles published daily. Therefore, an efficient and effective approach is needed to model topics related to the 2024 General Election from secret news articles.

Topic modelling is one of the techniques used in Natural Language Processing (NLP) to analyze text [7]. Topic modelling algorithms identify hidden patterns in a set of words distributed in a document collection. This technique

DOI : https://doi.org/10.25139/inform.v9i2.8429

results in a set of topics consisting of several groups of words that appear together in documents based on certain patterns. This research will use meta-description as a measurement. Meta Description is a description used to explain the purpose of the content created [8].

In previous research comparing several Natural Language Processing methods, Bidirectional Encoder Representations Transformers (BERT) was considered better than other models, including LDA [9]. BERT is a new sad language representation model designed to perform bidirectional representation training of unlabeled text [10]. One of the libraries that utilize BERT as a model for topic modelling is BERTopic [11]. BERTopic is a BERT transformer model specializing in topic modelling using the c-TF-IDF approach. This way, topic modelling can be interpreted easily without reducing keywords in the description of words used in topic modelling [12]. Therefore, this research aims to apply the BERTopic method in modelling topics in the meta description of the 2024 Election news articles and identifying, analyzing, and grouping the main topics that dominate the 2024 Election news coverage on the Detik.com news portal. In addition, this research also provides information on the trend of 2024 election topics in news articles on the portal.

Various visualizations will be employed to effectively present the results of this topic modelling research, including Distance Maps, Hierarchical Clustering, Term Score Decline, Top Word Scores, and Word Clouds. These visualizations aim to comprehensively understand the relationships and distributions of the topics identified. Distance Maps and Hierarchical Clustering will illustrate the proximity and hierarchical structure of the topics, Term Score Decline will show the relevance and significance of terms over time, Top Word Scores will highlight the most significant words in each topic, and Word Clouds will offer an intuitive visual representation of the key terms within each topic. These visualizations will help to interpret and analyze the dominant themes and trends in the 2024 Election news coverage on Detik.com. Additionally, this research aims to provide insights into the trends of 2024 election topics in news articles on the portal.

II. Literature Review

A. Topic Modeling

Topic modelling is an approach in text analysis aimed at identifying and extracting hidden topics or themes within a collection of documents. The primary goal of topic modelling is categorizing documents into specific topics based on their content [13].

B. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language model developed by Google that uses the Transformer architecture to understand the context of words in a text. BERT is designed to produce highly contextual word representations and can be utilized in various natural language processing (NLP) tasks [14].

C. Text Preprocessing

Preprocessing removes noise and normalizes word forms to make the vocabulary more uniform, thus reducing vocabulary volume [15]. According to Tu, text preprocessing aims to transform unstructured data into a more structured format [16]. The types of text preprocessing that can be implemented include [17][18].

1) Case folding: Case folding refers to converting all letters in the text to either lowercase or uppercase, depending on the needs of the analysis. This reduces variations due to case differences in the same words. For example, "Data", "data", and "DATA" will be converted to "data".

2) Cleaning: The cleaning stage involves removing irrelevant or disruptive characters or elements from the text. This may include removing punctuation marks, special characters, numbers, or unwanted web links.

3) Tokenizing: Tokenizing or tokenization divides the text into smaller units called tokens. Tokens can be words, phrases, or symbols, depending on the rules used. The result of tokenization will produce a set of tokens that can serve as a basis for further text analysis, such as classification, sentiment analysis, or language modelling.

4) Stopword Removal: Stopword removal eliminates common words, such as conjunctions and auxiliary words, that are considered unimportant in text data processing. This removal can reduce the data's dimensionality, thereby decreasing the data processing time for modelling.

D. BERTopic

BERTopic is a library that applies the BERT model for a specific purpose, namely topic modelling [19]. In topic modelling, BERTopic goes through 5 main processes plus one optional process, as shown in Figure 1 [20].

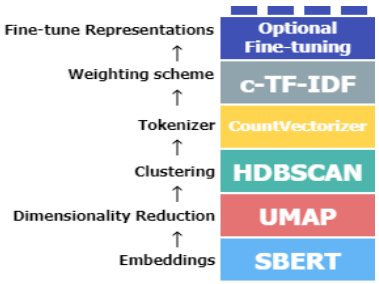

Figure 1. BERTopic process

1. Embeddings: BERTopic begins by transforming our input documents into numerical representations. Although there are many ways to achieve this, BERTopic uses the sentence transformers "all-MiniLM-L6-v2" for English and "paraphrase-multilingual-MiniLM-L12-v2" for multilingual because they can capture the semantic similarities between documents.

2. Dimensionality Reduction: One of the critical processes in BERTopic is the reduction of the dimensionality of the

DOI : https://doi.org/10.25139/inform.v9i2.8429

input embeddings. High-dimensional embeddings can often complicate clustering or topic grouping. UMAP (Uniform Manifold Approximation and Projection) is the standard in BERTopic because it can capture local and global high-dimensional space into lower dimensions.

3. Clustering: In BERTopic, the standard method used is HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) because it can capture structures with varying densities. However, in the BERTopic library, other clustering models can be adjusted as needed, such as K-means, Agglomerative Clustering, cuML HDBSCAN, and others.

4. Tokenizer: In topic modelling, the quality of topic representation is crucial for interpreting topics, conveying results, and understanding patterns. In BERTopic, the standard method used is CountVectorizer.

5. Weighting Scheme: In BERTopic, weighting is done using the c-TF-IDF method to obtain an accurate topic representation from the bag-of-words matrix. C-TF-IDF considers what makes documents within a cluster different from documents in other clusters. The c-TF-IDF equation is shown in Equation (1). Where the variable $tf_{X,C}$ is the term frequency of word X in class X, the f_X variable is the frequency of word X in all classes, and the A variable is the average number of words per class. Examples of writing equations can be seen in Equation (1).

$$W_{X,C} = |tf_{X,C}| \log(1 + \frac{A}{f_X}) \quad (1)$$

E. Evaluation

1. Topic Coherence: Topic coherence is a process that takes topics and a reference corpus as its input and then produces the coherence value of the topics as its output. The stages of this process include segmentation, probability calculation, confirmation measurement, and aggregation [21]. The range of topic coherence typically ranges from -1 to 1. Higher values indicate more coherent topics, while lower values indicate less coherent topics.

2. Silhouette Coefficient: The Silhouette Coefficient value can describe the quality of the generated clusters [22]. This value calculates the distance of data within a cluster or to its neighbor clusters [23]. It can be determined whether the model to be applied is suitable. Generally, the interpretation of the Silhouette Coefficient evaluation results can be explained in Table I [24].

TABLE I
INTERPRETATION OF SILHOUETTE COEFFICIENT RESULTS

Silhouette Coefficient Value	Interpretation
0.71 – 1.00	Strong structure
0.51 – 0.70	Good structure
0.26 – 0.50	Weak structure
≤ 0.25	Unstructured

3. Elbow: The Elbow Method is one of the evaluation methods used to analyze the performance of clustering results, usually in non-hierarchical clustering (partitioning) such as K-means [25]. The Elbow Method determines the optimal number of clusters in clustering by paying attention to the Sum of Squared Error (SSE) [26].

4. Davies-Bouldin Index: The Davies-Bouldin Index (DBI), often called the classification reliability index, is an internal cluster evaluation scheme that assesses the success of clustering based on the quality and compactness of the resulting clusters [27]. The lower the DBI value (approaching 0), the better the separation between clusters and the more compact the clusters [28].

5. Calinski-Harabasz Index: The Calinski-Harabasz Index (CHI) is a metric that calculates the ratio between the dispersion between clusters and the dispersion within clusters, measured as the sum of squared distances, for all clusters, or measures the ratio between within-cluster variance and between-cluster variance [29]. A high index value indicates that the data clusters are more separated from each other [30].

6. Dunn Index: The Dunn Index (DI) is a metric that measures the ratio between the minimum distance between two clusters and the maximum distance within the data clusters. A high index value indicates that the data clusters are more distinctly separated, and the results are better [31].

III. RESEARCH METHODOLOGY

The method in this research is qualitative and uses Topic Modeling techniques, using the BERTopic model, to analyze topics in online news articles related to the 2024 Indonesian General Election (Pemilu). BERTopic, as a Natural Language Processing (NLP) algorithm. In research on implementing the BERT (Bidirectional Encoder Representations Transformers) method, the object under study is news articles related to the 2024 election on news portals. This data collection is done through meta-description data of online news articles, in the case of this research, namely on the Detik.com news portal. The data used is the result of scraping the meta description of news articles in the election sub-channel from September 1, 2023, to February 14, 2024.

A. Data Collection

Data collection in this study was carried out using the scraping method on the metadata of news articles by utilizing the BeautifulSoup library. The scraping process is divided into four main stages to ensure the data obtained is complete and well-organized. The first stage is linked index scraping, where news article links are collected from the Detik.com news portal. The second stage involves the first stage of scraping, which is the collection of the initial metadata of each article obtained from the link index. In the third stage, the second stage, scraping is performed to complete the metadata that may not have been collected in the previous stage. The last stage is the merging of scraping data, where all data that has been collected is combined into one complete dataset. After

DOI : https://doi.org/10.25139/inform.v9i2.8429

the scraping process, the data obtained is still in raw form and requires further processing in the model implementation stage.

B. Preprocessing Data

The next stage is data preprocessing, where the raw data that has been collected is processed into clean and structured data according to the needs of the model implementation. This preprocessing process consists of several steps, including case folding to convert all text into lowercase letters, cleaning to remove unnecessary characters, tokenizing to break the text into word tokens, and stopword removal to remove common words in the data.

C. Model Implementation

Once the data was ready, the BERT model was chosen as the main model for modelling the 2024 Election topics due to its ability to produce embedding representations with optimal dimensions without requiring additional dimensionality reduction. BERTopic, based on BERT, was used for document clustering and keyword extraction. The results of topic modelling using BERTopic resulted in an uncertain number of topics due to the use of the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). It is a hierarchical clustering model as the default, so the number of clusters generated cannot be determined. The model also produces a topic "-1" referring to all outliers and is usually ignored [14]. This research will use another K-means model to avoid producing "-1" topics or outliers. This is a non-hierarchical model with several clusters based on Elbow's evaluation [15]. The range of the elbows that will be used is the value of k from 2 to 10.

D. Model Evaluation

After the model is implemented, an evaluation step is performed to measure the accuracy of topic modelling. This evaluation uses methods such as Elbow, Silhouette Coefficient, Davies Bouldin Index, Calinski Harabasz Index, and Dunn Index for clustering models and the topic coherence method to evaluate topic modelling results. The greater the value of topic coherence (approaching 1), the better the association value of the topic modelling results.

E. Deployment

The last process is deployment, where all the processes carried out on topic modelling using BERTopic are packaged more simply so that the results obtained provide informative and interesting data in the form of a data dashboard system. The dashboard can identify topics that dominate the current conversation around the 2024 Election on various online news portals. The results will be implemented on a simple system in the form of a website that provides information on the model implementation results and the dominating news trends at any given time. The system in this research was created using the Streamlit Framework in the Python programming language.

IV. RESULT AND DISCUSSION

A. Dataset Collection Implementation

1. Scraping link index: The number of links that can be retrieved is 827 data. The results of the link index scraping stage are page URLs such as https://news.detik.com/pemilu/indeks/1?date=09/01/2023 to https://news.detik.com/pemilu/indeks/20?date=02/14/2024.

2. First Scraping: The scraping results are in the title, date, link, and metadata of each news item in the year range of 10,519 data from September 1, 2023, to January 10, 2024. The first stage of scraping results is shown in the first stage of scraping news article data related to the 2024 Election on the Detik.com news portal. Several articles were successfully collected. One of them is an article entitled published on September 1, 2023, with the URL https://news.detik.com/pemilu/d-6908814/eks-tim-8-cerita-sempat-bahas-cak-imin-jadi-cawapres-anies-di-desember. Another article that was successfully collected was entitled "*Cuti dari Menhan, Prabowo Kampanye di 3 Provinsi dalam Sehari di Sumatera*," published on January 10, 2024, with URL *https://news.detik.com/pemilu/d-7133349/cuti-dari-menhan-prabowo-kampanye-di-3-provinsi-dalam-sehari-di-sumatera*

3. Second Scraping: Same as the first step, scraping results, in the second step, the scraping results are in the form of title, date, link, and metadata of each news in the year range with a total of 4,500 data.

4. Scraping Data Merge: The last step is merging the results of the first and second scraping steps. The total period is from September 1, 2023, to February 14, 2024, with the number of articles obtained totalling 15,019 news articles. Then, the scraping results are stored as .csv files, which will be processed at the data preprocessing step. The results of the dataset collection are shown in Table II.

TABLE II
DATASET COLLECTION RESULT

Title	Date	Link	Description
Komunitas Sepak Bola Banten Juga Ikut Gaspoll Bro, Siap Dukung Prabowo-Gibran	11/1/2024	*https://news.detik.com/pemilu/d-7137213/komunitas-sepak-bola-banten-juga-ikut-gaspoll-bro-siap-dukung-prabowo-gibran*	*Komunitas Sepak Bola Banten juga ikut bergabung bersama relawan Gaspoll Bro. Langkah itu dilakukan agar dapat memenangkan Prabowo-Gibran di Pilpres 2024.*
.....	.....	.....	.....
Momen Mahfud Didoakan Menang oleh Cucu di Belanda Lewat Zoom	14/02/24	*https://news.detik.com/pemilu/d-7191736/momen-mahfud-didoakan-menang-oleh-cucu-di-belanda-lewat-zoom*	*Cawapres nomor urut 3, Mahfud Md mengisi masa tenang hari terakhir dengan zoom bersama cucunya, Irada yang berada di Belanda.*

B. Text Preprocessing Implementation

Text preprocessing is implemented on the scraped dataset, as shown in Table II, for the description column data so that another column can be deleted. Text preprocessing uses four processes: case folding, cleaning, tokenizing, and stopword removal.

DOI : https://doi.org/10.25139/inform.v9i2.8429

1. Case folding: Case folding converts all characters in the data into lowercase letters. The process results before and after case folding are shown in Table III.

TABLE III
CASE-FOLDING RESULT

Before Case Folding	After Case Folding
Soal Cak Imin jadi Cawapres Anies sempat dibahas dalam rapat Tim 8 pada Desember 2022. Eks Tim 8 menyebut manuver Cak Imin itu tak mengejutkan.	*soal cak imin jadi cawapres anies sempat dibahas dalam rapat tim 8 pada desember 2022. eks tim 8 menyebut manuver cak imin itu tak mengejutkan.*
.....	.....
Cawapres nomor urut 3, Mahfud Md mengisi masa tenang hari terakhir dengan zoom bersama cucunya, Irada yang berada di Belanda.	*cawapres nomor urut 3, mahfud md mengisi masa tenang hari terakhir dengan zoom bersama cucunya, irada yang berada di belanda.*

*2. Cleaning: Cle*aning is removing punctuation from characters in a dataset. The results of this process, both before and after cleaning, are shown in Table IV.

TABLE IV
CLEANING RESULT

Before Cleaning	After Cleaning
soal cak imin jadi cawapres anies sempat dibahas dalam rapat tim 8 pada desember 2022. eks tim 8 menyebut manuver cak imin itu tak mengejutkan.	*soal cak imin jadi cawapres anies sempat dibahas dalam rapat tim 8 pada desember 2022 eks tim 8 menyebut manuver cak imin itu tak mengejutkan*
.....	.....
cawapres nomor urut 3, mahfud md mengisi masa tenang hari terakhir dengan zoom bersama cucunya, irada yang berada di belanda.	*cawapres nomor urut 3 mahfud md mengisi masa tenang hari terakhir dengan zoom bersama cucunya irada yang berada di belanda*

3. Tokenizing: Tokenizing breaks the text data in the dataset into word tokens by removing spaces in the text. The results of the process before and after tokenizing are shown in Table V.

TABLE V
TOKENIZING RESULT

Before Tokenizing	After Tokenizing
soal cak imin jadi cawapres anies sempat dibahas dalam rapat tim 8 pada desember 2022 eks tim 8 menyebut manuver cak imin itu tak mengejutkan	*['soal', 'cak', 'imin', 'jadi', 'cawapres', 'anies', 'sempat', 'dibahas', 'dalam', 'rapat', 'tim', '8', 'pada', 'desember', '2022', 'eks', 'tim', '8', 'menyebut', 'manuver', 'cak', 'imin', 'itu', 'tak', 'mengejutkan']*
.....	.....
cawapres nomor urut 3 mahfud md mengisi masa tenang hari terakhir dengan zoom bersama cucunya irada yang berada di belanda	*['cawapres', 'nomor', 'urut', '3', 'mahfud', 'md', 'mengisi', 'masa', 'tenang', 'hari', 'terakhir', 'dengan', 'zoom', 'bersama', 'cucunya', 'irada', 'yang', 'berada', 'di', 'belanda']*

4. Stopword removal: Stopword removal is removing common words from the dataset. The results of the process before and after stopword removal are shown in Table VI.

TABLE VI
STOPWORD REMOVAL RESULT

Before Stopword Removal	After Stopword Removal
['soal', 'cak', 'imin', 'jadi', 'cawapres', 'anies', 'sempat', 'dibahas', 'dalam', 'rapat', 'tim', '8', 'pada', 'desember', '2022', 'eks', 'tim', '8', 'menyebut', 'manuver', 'cak', 'imin', 'itu', 'tak', 'mengejutkan']	*['cak', 'imin', 'cawapres', 'anies', 'dibahas', 'rapat', 'tim', '8', 'desember', '2022', 'eks', 'tim', '8', 'menyebut', 'manuver', 'cak', 'imin', 'mengejutkan']*
.....	.....
['cawapres', 'nomor', 'urut', '3', 'mahfud', 'md', 'mengisi', 'masa', 'tenang', 'hari', 'terakhir', 'dengan', 'zoom', 'bersama', 'cucunya', 'irada', 'yang', 'berada', 'di', 'belanda']	*['cawapres', 'nomor', 'urut', '3', 'mahfud', 'md', 'mengisi', 'tenang', 'zoom', 'cucunya', 'irada', 'belanda']*

After all the text preprocessing processes are completed, the results of the text preprocessing are returned in the form of whole sentences as postprocessing datasets, as shown in Table VII.

TABLE VII
POSTPROCESSING RESULT

Dataset Postprocessing
cak imin cawapres anies dibahas rapat tim 8 desember 2022 eks tim 8 menyebut manuver cak imin mengejutkan
.....
cawapres nomor urut 3 mahfud md mengisi tenang zoom cucunya irada belanda

C. BERTopic Modeling Implementation

Topic modelling using BERTopic in this study goes through several main processes, as shown in Figure 1, related to the BERTopic process: embeddings, dimensionality reduction, clustering, tokenizer, and weighting scheme. Researchers modified the embedding process using "distiluse-base-multilingual-cased-v1" and clustering using the K-Means model to avoid outliers in the topic modelling results.

1. Embeddings: This process is performed using sentence-transformers "distiluse-base-multilingual-cased-v1". This process generates an array of each document as a numerical representation with high array dimensions. 'Document' refers to each of the 15,019 news articles obtained from the scraping process that has undergone preprocessing (data postprocessing), as shown in Table VIII. Each array represents an individual news article.

TABLE VIII
EMBEDDINGS RESULT

Before Embeddings	After Embeddings
cak imin cawapres anies dibahas rapat tim 8 desember 2022 eks tim 8 menyebut manuver cak imin mengejutkan	[0.01337722409516573, 0.0305610541254282, 0.008294962346553802, 0.050004929304122925, 0.018364250659942627,..., 0.00913078710436821]
.....	.....
cawapres nomor urut 3 mahfud md mengisi tenang zoom cucunya irada belanda	[0.05712178722023964, 0.017337003722786903, 0.0281606987118721, 0.03628504276275635, 0.04624389111995697,…, 0.07578016072511673]

2. Dimensionality Reduction: This process is done using UMAP with the same approach as BERTopic, namely UMAP dimensional reduction parameters, namely with parameters

DOI : https://doi.org/10.25139/inform.v9i2.8429

n_neighbors=15 (number of nearest neighbors), n_components=5 (number of mapped dimensions), min_dist=0.0 (minimum distance between points), and metric='cosine' (distance measurement metric). In simple terms, these parameters control how UMAP reduces data dimensionality concerning the neighbor structure, the number of resulting dimensions, the minimum distance between points, and the similarity metric between data points. The results before and after performing dimensionality reduction are shown in Table IX.

TABLE IX
DIMENSIONALITY REDUCTION RESULT

Before Dimensionality Reduction	After Dimensionality Reduction
[0.01337722409516573, 0.0305610541254282, 0.008294962346553802, 0.050004929304122925, 0.018364250659942627, …, 0.00913078710436821]	[-0.2608505189418793, 9.72555160522461, 9.114203453063965, 10.026761054992676, 6.716017246246338]
.....	.....
[0.05712178722023964, 0.017337003722786903, 0.0281606987118721, 0.03628504276275635, 0.04624389111995697, …, 0.07578016072511673]	[8.69973087310791, 8.878094673156738, 4.56990909576416, 8.575509071350098, 5.26642370223999]

3. Clustering: The clustering process is carried out to group documents into the same groups using documents subjected to a dimensionality reduction process. In this research, the clustering model used is K-Means with the same number of clusters or k value according to the results of the Elbow evaluation. The evaluation results using the Elbow method are shown in Figure 2.

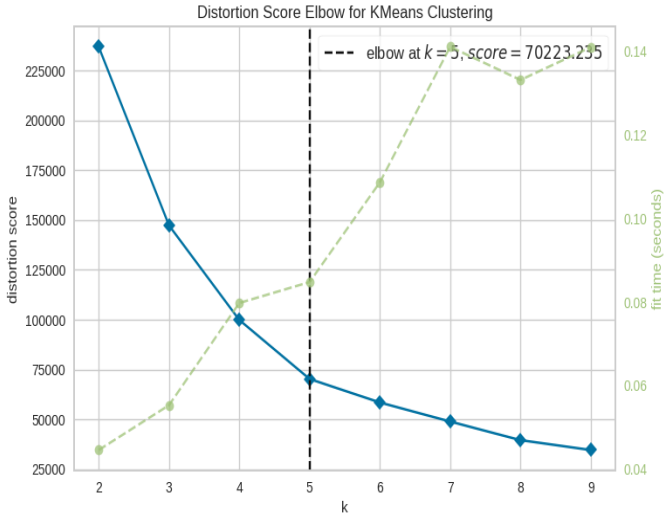

Figure 2. Elbow evaluation result

In several experiments, Elbow's evaluation results showed the optimal k value between 4 and 5 with Distortion Score (SSE) = 70223.257 with the value of *k=5* because a fixed *k* value was needed for implementation in the system, and by using the k value, the Silhouette Score result of 0.566 was obtained. The resulting silhouette plot graph is shown in Figure 3.

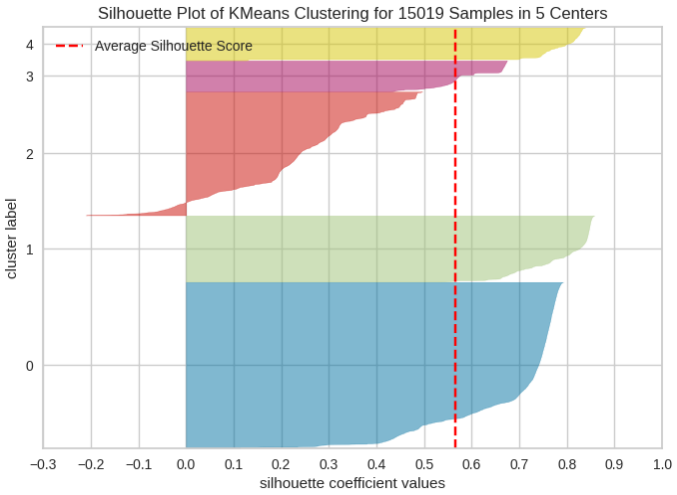

Figure 3. Graphics of silhouette plot

Based on the Silhouette Score and Silhouette plot results, the K-Means model clustering with a value of k equal to 5 produces good clustering with clear inter-cluster distances. The SSE value of 70223.257 evaluations provides an overview of the cluster distribution. Davies-Bouldin Index 0.758 shows the cluster has a relatively good level of inter-cluster separation with good closeness within the cluster, Calinski-Harabasz Index 20083.489 shows good compact inter-cluster separation, Dunn Index 0.003 shows outliers that cause overlapping clusters and lack of clear separation.

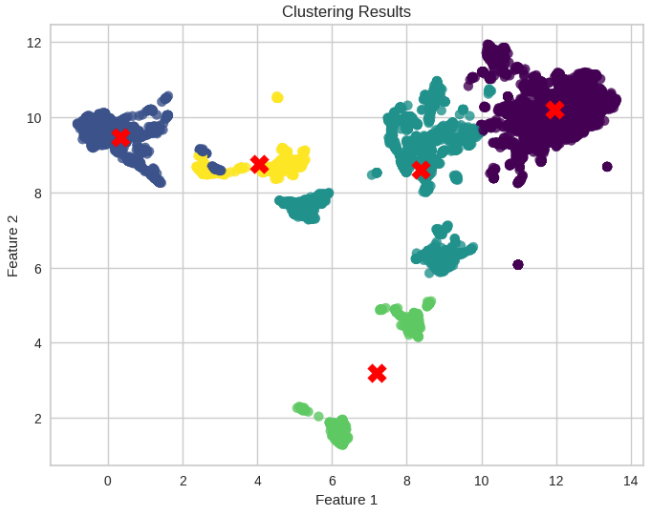

Figure 4. Clustering Result

The evaluation results show that implementing the K-Means model with a value of k equal to 5 again emphasizes if the clustering results are good. To create a two-dimensional visualization of the clustering results, the author uses 'Feature 1' and 'Feature 2' from each document that has been reduced to 5 features using UMAP. The clustering results in two dimensions are shown in Figure 4.

4. Tokenizer: The tokenizer process creates a topic representation using CountVectorizer. This process results in

DOI : https://doi.org/10.25139/inform.v9i2.8429

the frequency of occurrence of each word in each document. The rows in the table represent documents or text data in the context of this research meta-description on news articles. The columns represent features or words in all data, and the value in each table cell is the frequency of words' occurrence in related documents. The tokenizer results using CountVectorizer are shown in Table X.

TABLE X
TOKENIZER RESULT

	0	1	7	...	zulkarnain	zulkieflimansyah	zulkifli
0	0	0	0	...	0	0	0
1	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...
15017	0	0	0	...	0	0	0
15018	0	0	0	...	0	0	0

5. Weighting Scheme: The weighting scheme process is the creation of topic representations using the c-TF-IDF method. This process results in topic modelling, which involves the number of topics according to the number of clusters determined at the beginning. Based on the weighting scheme process results, topic 0 is the dominating topic in the dataset, with a total of 5.215 news articles. It can be concluded that topics related to presidential and vice-presidential candidates are often reported by the Detik.com news portal in the election sub-channel. The results of the weighting scheme using the c-TF-IDF method are shown in Table XI.

TABLE XI
WEIGHTING SCHEME RESULT

Topic	Count	Name	Representation	Representative_ Docs
0	5215	*0_ganjar_prabowo_anies_imin*	*['ganjar', 'prabowo', 'anies', 'imin', 'ketua', 'cak', 'gibran', 'pdip', 'mahfud', 'prabowogibran']*	*['ketum psi kaesang pangarep bertemu ketum partai gerindra prabowo subianto .....*
1	3191	*1_2024_pemilu_pilpres_suara*	*['2024', 'pemilu', 'pilpres', 'suara', 'survei', 'kpu', 'hasil', 'prabowo', 'ganjar', 'anies']*	*['elektabilitas pdip unggul hasil survei terbaru poltracking menjelang .....*
2	2604	*2_presiden_jokowi_demokrat_politik*	*['presiden', 'jokowi', 'demokrat', 'politik', 'partai', 'pdip', 'prabowo', 'ganjar', 'calon', 'ketua']*	*['calon presiden nomor urut 3 ganjar pranowo skor 5 terkait penegakan hukum era presiden joko widodo jokowi', ...*
3	2046	*3_debat_cawapres_indonesia_capres*	*['debat', 'cawapres', 'indonesia', 'capres', 'pilpres', '2024', 'nomor', 'anies', 'urut', 'kpu']*	*['debat capres ketiga malam calon wakil presiden nomor urut 2 gibran rakabuming raka prabowo .....*
4	1963	*4_raka_rakabuming_nomor_urut*	*['raka', 'rakabuming', 'nomor', 'urut', 'gibran', 'prabowo', 'program', 'capres', 'ganjar', 'subianto']*	*['pasangan calon presidenwakil presiden nomor urut 2 prabowo subianto gibran rakabuming raka menghadiri .....*

D. BERTopic Modeling Evaluation

In addition to providing topic representation results, topic modelling using BERTopic produces several graphics or visualizations such as intertropical distance map, hierarchical clustering, topic word scores, similarity matrix, and term score decline per topic and the addition of word cloud graphics by the author and evaluation using the topic coherence method. The following is the graphical result of topic modelling using BERTopic.

1. Intertopic Distance Map: An intertopic distance map is a graphic that shows how close or far different topics are in the topic modelling results—a graphical intertropical distance map of topic modelling results using BERTopic as follows. Using the graphical intertopic distance map of the topic modelling results using BERTopic, topics with similarities can be grouped into topics 0, 2, and 3 as group A and topics 1 and 4 as group B. Then, the topic groups have significant differences. The intertopic distance map graph is shown in Figure 5.

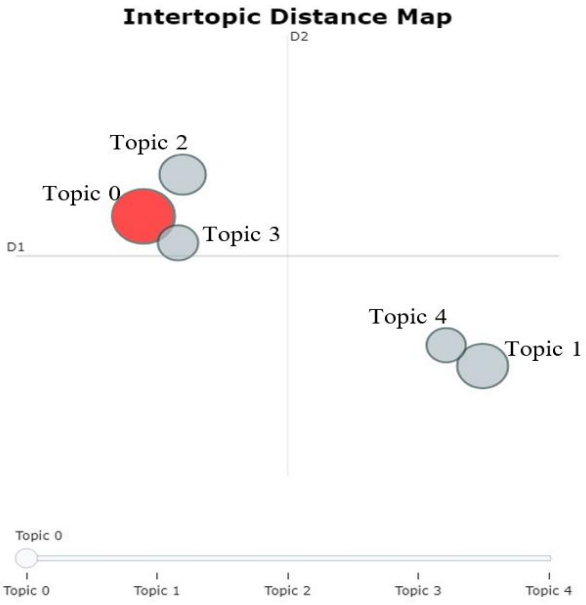

Figure 5. Graphic of Intertopic Distance Map

2. Hierarchical clustering: Hierarchical clustering is a graph that shows the relationship between the resulting topics. The graph determined which topics should be combined (have similar properties) and which should be separated (have different properties). Based on the hierarchical clustering graph, for example, if the researcher wants to reduce it to 3

DOI : https://doi.org/10.25139/inform.v9i2.8429

topics, it can be done by combining topics 0 and 2 and topics 4 and 3. The hierarchical clustering graph is shown in Figure 6.

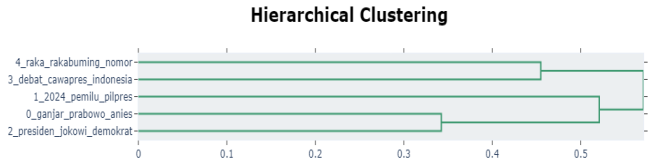

Figure 6. Graphic of Hierarchical Clustering

3. Term Score Decline Per Topic: Term score decline per topic is a graph that refers to a keyword score decline that occurs specifically in the context of a particular topic. This makes it possible to see how the relevance of a keyword changes or decreases with further research on that topic. The term score decline per topic graph is shown in Figure 7.

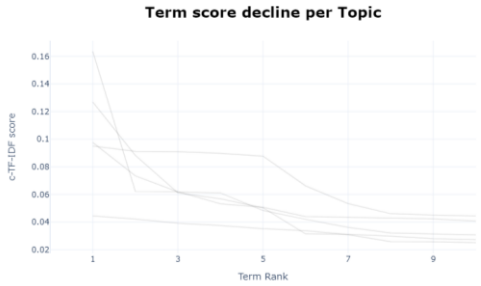

Figure 7. Graphic of Term Score Decline Per Topic

4. Similarity matrix: A similarity matrix is a graphic that shows the level of similarity between topics in the form of a heatmap. The graphic similarity matrix of topic modelling results using BERTopic is as follows. Based on the graphical similarity matrix of the topic modelling results using BERTopic, there are no topics with a high level of similarity. It can be concluded that the topic modelling results are good. The similarity matrix graph is shown in Figure 8.

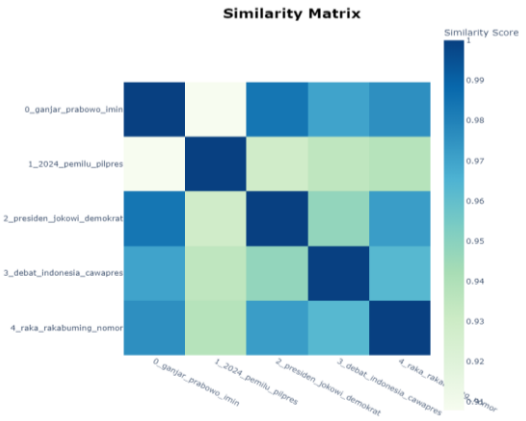

Figure 8. Graphic of Similarity Matrix

5. Topic word scores: Topic word scores are graphs that show the frequency of each word in the topic representation for each cluster or related topic. Using graphical topic word scores, the results of topic modelling using BERTopic in Figure 4.8, for example, in topic one, the most frequently occurring words are '2024' and 'election'. The topic word scores graph is shown in Figure 9.

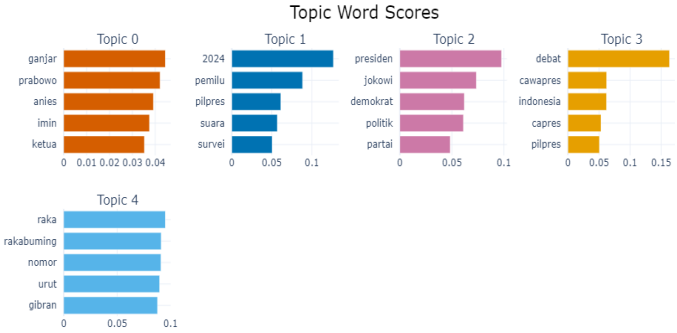

Figure 9. Graphics of Topic Word Scores

6. Word cloud: Different topic word scores with word clouds are graphics that show the frequency of each word in the entire dataset. Larger words indicate that the word has a high frequency or appears frequently in the entire dataset. Based on the word cloud graphic, the words that often appear in the dataset obtained from the meta description on the election sub-channel on Detik.com are the names of presidential and vice-presidential candidates such as *'ganjar'*, *'prabowo'*, *'anies'*, and others. The word cloud graph is shown in Figure 10.

Figure 10. Word cloud

7. Topic coherence: Topic coherence is an evaluation method to measure whether the chosen words reflect a specific topic. The results of the topic coherence evaluation using BERTopic, as shown in Table XII, indicate an average topic coherence of 0.0902.

TABLE XII
WEIGHTING SCHEME RESULT

Topic	Count	Coherence Value
0	5215	0.101180854
1	3191	0.091238307
2	2604	0.083697987
3	2046	0.086670586
4	1963	0.088206257
Average		**0.090198798**

DOI : https://doi.org/10.25139/inform.v9i2.8429

8. System Implementation: The system implementation is done using the Streamlit framework. Streamlit is an effective tool for building interactive and responsive user interfaces for data science applications, making it easy for users to run topic modelling without deep technical knowledge in coding. The system implementation aims to make research results easily accessible to users publicly. The home page provides information about the modelling topics in this study, such as dataset information for the model used, shown in Figure 11.

Figure 11. Home Page Implementation

Figure 12 The BERTopic page is a menu for modelling topics using previously described methods.

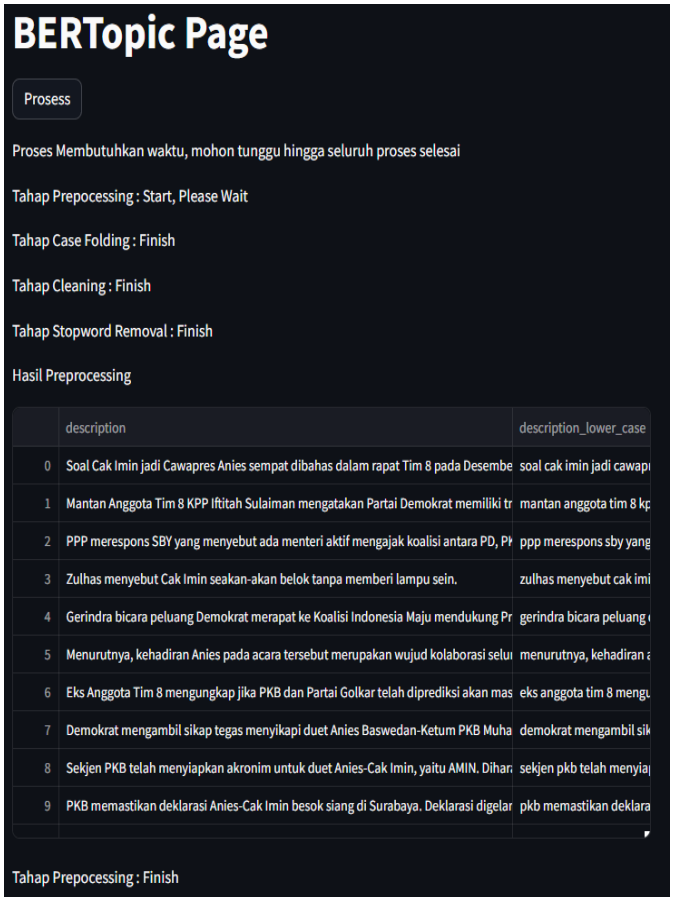

Figure 12. Bertopic Page Implementation

The article information menu provides information about the topic modelling result document using BERTopic, shown in Figure 13.

	Document	Topic	Name
0	soal cak imin jadi cawapres anies sempat dibahas dalam rapat tim 8 pada desember	1	1_di_gibran_rak
1	mantan anggota tim 8 kpp iftitah sulaiman mengatakan partai demokrat memiliki tra	1	1_di_gibran_rak
2	ppp merespons sby yang menyebut ada menteri aktif mengajak koalisi antara pd pks	0	0_yang_di_dan_
3	zulhas menyebut cak imin seakanakan belok tanpa memberi lampu sein	0	0_yang_di_dan_
4	gerindra bicara peluang demokrat merapat ke koalisi indonesia maju mendukung pra	2	2_presiden_joko
5	menurutnya kehadiran anies pada acara tersebut merupakan wujud kolaborasi selur	0	0_yang_di_dan_
6	eks anggota tim 8 mengungkap jika pkb dan partai golkar telah diprediksi akan masu	1	1_di_gibran_rak
7	demokrat mengambil sikap tegas menyikapi duet anies baswedanketum pkb muhair	2	2_presiden_joko
8	sekjen pkb telah menyiapkan akronim untuk duet aniescak imin yaitu amin diharap a	0	0_yang_di_dan_
9	pkb memastikan deklarasi aniescak imin besok siang di surabaya deklarasi digelar di	0	0_yang_di_dan_

Figure 13. Article Information Page Implementation

V. CONCLUSION

Implementing BERTopic for election topic modelling was done using meta descriptions of news articles related to the 2024 election. The dataset, obtained through web scraping from September 1, 2023, to February 14, 2024, consisted of 15,019 articles. The text preprocessing included four stages: case folding, cleaning, tokenizing, and stopword removal to ensure the data was clean and structured. The topic modelling using BERTopic involved five stages: embeddings using sentence-transformers "distiluse-base-multilingual-cased-v1", dimensionality reduction using UMAP, clustering with K-Means (with an optimal k value of 5 as determined by the Elbow method), tokenizing using CountVectorizer, and weighting using c-TF-IDF. The topic coherence score achieved was 0.0902.

This study successfully identified five main topics that dominated the news about the 2024 election on detik.com, including a focus on presidential and vice-presidential candidates, highlighting significant attention to key figures in the election, represented by keywords like *'ganjar'*, *'prabowo'*, *'anies'*, and *'imin'* across 5,215 articles; General election and presidential election topics, including related votes and surveys, represented by keywords such as '2024', *'pemilu'*, *'pilpres'*, and *'suara'* across 3,191 articles; News about Joko Widodo President and his political dynamics, represented by keywords like *'presiden'*, *'jokowi'*, *'demokrat'*, and *'politik'* across 2,604 articles; News about presidential and vice-presidential debates, represented by keywords like *'debat'*, *'cawapres'*, *'indonesia'*, and *'capres'* across 2,046 articles; Focus on Gibran Rakabuming Raka and issues related to him, represented by keywords like *'raka'*, *'rakabuming'*, *'nomor'*, and *'urut'* across 1,963 articles. From the topic modelling results, the main topics dominating the news coverage of the 2024 election on detik.com were identified. Among these, the topic related to presidential and vice-presidential candidates was the most frequently mentioned.

Based on the findings of this research, several suggestions for future studies are proposed: Comparing the BERTopic model with other topic modelling methods such as Correlated Topic Models (CTM), Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), Non-Negative Matrix Factorization

DOI : https://doi.org/10.25139/inform.v9i2.8429

(NMF), among others; Utilizing different embedding methods; Expanding the dataset to include news articles from other news portals related to the 2024 election; Comparing multiple news portals to determine which topics dominate each portal's coverage of the 2024 election.

REFERENCES

[1] K. K. Sabat, "Kepemimpinan Ideal Bagi Generasi Milenial," HARVESTER: Jurnal Teologi dan Kepemimpinan Kristen, vol. 6, no. 2, pp. 149–159, 2021, doi: 10.52104/harvester.v6i2.59.

[2] F. I. R. Firamadhina and H. Krisnani, "Perilaku Generasi Z Terhadap Penggunaan Media Sosial Tiktok: TikTok Sebagai Media Edukasi dan Aktivisme," Share: Social Work Journal, vol. 10, no. 2, pp. 199–208, 2021, doi: 10.24198/share.v10i2.31443.

[3] D. M. Solikha and H. P. Purba, "Perbedaan Value Pada Generasi X dan Y di Indonesia," Jurnal Diversita, vol. 8, no. 1, pp. 38–43, 2022, doi: 10.31289/diversita.v8i1.5188.

[4] Y. S. Putra, "Theoritical Review: Teori Perbedaan Generasi," Jurnal STIE AMA, no. 1952, pp. 123–134, 2017.

[5] E. Nur, "Peran Media Massa dalam Menghadapi Serbuan Media Online," Majalah Semi Ilmiah Populer Komunikasi Massa, vol. 2, no. 1, pp. 51–64, 2021.

[6] Detikcom, "Detikcom Company Profile," Detikcom. [Online]. Available: https://detiknetwork.com/logo/logo/pdf-Company-Profile-detikcom-2021.pdf

[7] C. C. Aggarwal and C. Zha, Mining Text Data. Springer publishing company, 2012.

[8] P. Septiani and H. Kurniawan, "Analisa Penggunaan Keyword Untuk Implementasi Search," Jurnal Teknologi Informasi, vol. 15, no. 3, pp. 83–91, 2020.

[9] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts.," Frontiers in sociology, vol. 7, p. 886498, 2022, doi: 10.3389/fsoc.2022.886498.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[11] G. Malki, "Efficient Sentiment Analysis and Topic Modeling in NLP using Knowledge Distillation and Transfer Learning," School of Electrical Engineering and Computer Science, 2023.

[12] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv, 2022.

[13] Y. Matira and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," Estimasi: Journal of Statistics and Its Application, pp. 53–63, 2023.

[14] B. Kurniawan, A. A. Aldino, and A. R. Isnain, "Sentimen Analisis Terhadap Kebijakan Penyelenggara Sistem Elektronik (Pse) Menggunakan Algoritma Bidirectional Encoder Representations from Transformers (Bert)," J. Teknol. dan Sist. Inf, vol. 3, no. 4, pp. 98–106, 2022.

[15] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," J-SAKTI (Jurnal Sains Komputer Dan Informatika), vol. 5, no. 2, pp. 697–711, 2021.

[16] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," Jurnal Informatika dan Rekayasa Perangkat Lunak, vol. 4, no. 1, pp. 74–80, 2023.

[17] M. Nashrullah and D. A. Efrilianda, "Sentiment Analysis of Independent Campus Policy on Twitter Using Support Vector Machine and Naïve Bayes Classifier," Journal of Advances in Information Systems and Technology, vol. 4, no. 1, pp. 13–23, 2022.

[18] A. Sujjada and A. Fergina, "Implementasi Metode Vector Space Model Untuk Deteksi Emosi Menggunakan Data Teks Twitter," Jurnal RESTIKOM: Riset Teknik Informatika dan Komputer, vol. 3, no. 3, pp. 116–129, 2021.

[19] G. Malki, "Efficient Sentiment Analysis and Topic Modeling in NLP using Knowledge Distillation and Transfer Learning." 2023.

[20] Maarten Grootendorst, "BERTopic." Accessed: March 27, 2024. [Online].

[21] I. M. A. Mahesastraa and I. D. M. B. A. Darmawana, "Pemodelan Topik Teks Berita Menggunakan DistilBERT," vol. 1, no. 1, 2022.

[22] R. N. Fahmi, M. Jajuli, and N. Sulistiyowati, "Analisis pemetaan tingkat kriminalitas di kabupaten Karawang menggunakan Algoritma K-Means," INTECOMS: Journal of Information Technology and Computer Science, vol. 4, no. 1, pp. 67–79, 2021.

[23] F. Nuraeni, D. Kurniadi, and G. F. Dermawan, "Pemetaan Karakteristik Mahasiswa Penerima Kartu Indonesia Pintar Kuliah (KIP-K) menggunakan Algoritma K-Means++," Jurnal Sisfokom (Sistem Informasi Dan Komputer), vol. 11, no. 3, pp. 437–443, 2023.

[24] D. E. Herwindiati and T. Handhayani, "Clustering Data Covid-19 Di Indonesia Menggunakan Intelligent K-Means," Jurnal Ilmu Komputer dan Sistem Informasi, vol. 10, no. 2, 2022.

[25] L. Qadrini, "Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017," J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika, vol. 13, no. 2, pp. 5–11, 2020.

[26] H. Fitriyah, E. M. Safitri, N. Muna, M. Khasanah, D. A. Aprilia, and D. Nurdiansyah, "Implementasi Algoritma Clustering Dengan Modifikasi Metode Elbow Untuk Mendukung Strategi Pemerataan Bantuan Sosial Di Kabupaten Bojonegoro," Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika, vol. 4, no. 3, pp. 1598–1607, 2023.

[27] N. Azmi, H. S. Hafsah, Y. Yuyun, and H. Hazriani, "Penerapan Metode K-Means Clustering Dalam Mengelompokkan Data Penjualan Obat pada Apotek M23," Prosiding SISFOTEK, vol. 7, no. 1, pp. 244–248, 2023.

[28] R. Dwirahmanto and A. Bisri, "Menentukan Nilai K Pada Metode K-Means Menggunakan Teknik Grid Search Untuk Strategi Produk Pakaian Medis," Jurnal Informatika Multi, vol. 1, no. 2, pp. 93–103, 2023.

[29] A. M. Sikana and A. W. Wijayanto, "Analisis Perbandingan Pengelompokan Indeks Pembangunan Manusia Indonesia Tahun 2019 dengan Metode Partitioning dan Hierarchical Clustering," J. Ilmu Komput, vol. 14, no. 2, pp. 66–78, 2021.

[30] D. A. Saidah, R. Santoso, and T. Widiharih, "Pengelompokan Provinsi Di Indonesia Berdasarkan Indikator Kesehatan Lingkungan Menggunakan Metode Partitioning Around Medoids Dengan Validasi Indeks Internal," Jurnal Gaussian, vol. 11, no. 2, pp. 302–312, 2022.

[31] Y. Matira and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," Estimasi: Journal of Statistics and Its Application, pp. 53–63, 2023.

[32] L. Qadrini, "Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017," Jurnal Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika, vol. 13, no. 2, pp. 5–11, 2020, doi: 10.36456/jstat.vol13.no2.a2886.

DOI : https://doi.org/10.25139/inform.v9i2.8429