e-ISSN: 2614-8404
p-ISSN:2776-3234

Comparison of LDA and BERTopic in Identifying Public Issues in the MBG Program

Nur Hayati[1], Saikin[2]*, Hairul Fahmi[3]
[1,2]Information Systems Study Program, STMIK Lombok, Praya, Indonesia
[3]Informatics Engineering Study Program, STMIK Lombok, Praya, Indonesia
Email: [1]nnurhayaati83@gmail.com, [2]eken.apache@gmail.com, [3]iroel.ami@gmail.com

Abstract −The Free Nutritious Food Program (MBG) is a government policy that has generated various public responses and opinions on social media. The large amount of unstructured text data. This study aims to compare the performance of the Latent Dirichlet Allocation (LDA) and BERTopic methods in identifying public issues related to the MBG program on TikTok data. The dataset used amounted to 13,538 data obtained through a scraping process based on keywords related to MBG. The research stages include text preprocessing, bigram and trigram formation, text representation using TF-IDF and embedding, topic modeling, and evaluation using coherence score and topic diversity. The results showed that the LDA method produced better evaluation performance with a coherence score of 0.5098 and a topic diversity of 0.9000. Meanwhile, BERTopic produced a coherence score of 0.4133 and a topic diversity of 0.7667, but was able to produce topics that were more contextual and semantically representative. Based on these results, LDA is superior in terms of the stability and quality of word associations between topics, while BERTopic is more effective in understanding the context of issues in short and unstructured social media data.

Keywords – Topic Modeling, LDA, BERTopic, Social Media, Free Nutritious Food.

I. INTRODUCTION

The Free Nutritious Food Program (MBG) is a government initiative aimed at improving the nutritional quality of the community and supporting the development of healthier and more productive human resources. As a public policy that directly impacts the community, the implementation of this program not only impacts health aspects but also triggers various responses, perceptions, and discussions among the public. Differences in social and economic backgrounds and levels of public understanding of the program have led to the emergence of diverse views, both supportive and critical. In this context, understanding how public opinion is formed and developed is crucial for policy evaluation. Along with the increasing use of digital platforms, various public responses to the MBG program are now widely recorded and disseminated through social media, thus opening up opportunities for further analysis as a rich source of information on emerging issues in society [1][2].

The large amount of data generated from social media related to the MBG program presents unique challenges in the analysis process, especially because the data is unstructured, dynamic, and contains various language variations. To overcome this, approaches based on text mining and natural language processing (NLP) are effective solutions in extracting important information from large collections of text. One technique widely used in this context is topic modeling, which aims to identify patterns of topics or key issues hidden within text data without requiring prior labels. Methods such as Latent Dirichlet Allocation (LDA) have been widely used due to their ability to cluster documents based on word distribution, although they still have limitations in capturing deeper semantic context, especially in short texts such as social media. As technology advances, more modern approaches such as BERTopic are being developed that utilize embedding-based representations, thus being able to produce more contextual and coherent topics. Therefore, selecting the right method is a crucial factor in producing accurate and relevant public issue analysis [1][3][4][5].

Although various topic modeling methods have been widely used to analyze text data, selecting the right model remains a significant challenge, especially in the context of short and unstructured social media data. Classical methods such as Latent Dirichlet Allocation (LDA) have proven effective in identifying latent topics in various types of documents, but tend to be less than optimal in capturing complex semantic relationships in short texts. On the other hand, newer approaches such as BERTopic offer advantages through the use of transformer-based embeddings that are able to produce more contextual and coherent topics. Several recent studies have shown differences in performance between the two methods, where BERTopic tends to excel in understanding context, while LDA still has advantages in interpretability and computational efficiency. However, the results of this comparison still depend on the characteristics of the data used, so further studies are needed to evaluate the performance of both methods in specific contexts, such as public issue analysis on social media data related to the MBG program [6][7][8][9][10].

Several previous studies have examined the application of topic modeling in text data analysis, particularly in the context of social media and digital documents. A study by Grootendorst (2022) introduced BERTopic as an embedding-based approach capable of generating topics with a higher degree of coherence than traditional methods [1]. Furthermore, Egger and Yu (2022) compared LDA and BERTopic on Twitter data, showing that BERTopic was superior in capturing semantic context, while LDA remained relevant in terms of model interpretability [6]. Another study by Bianchi et al. (2021) also emphasized that the use of contextualized embeddings can improve the quality of the generated topics [3]. On the other hand, a

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

study by Pratiwi and Tania (2025) showed that a combination of several topic modeling methods such as LDA, BERTopic, and NMF can provide more comprehensive insights in the information extraction process [7]. In addition, Liu (2024) revealed that the performance of each method is highly dependent on the characteristics of the data used, so further evaluation in different data contexts is needed [4]. However, most of these studies have not specifically examined the comparison of the two methods in the context of public issue analysis of government programs such as MBG, thus opening up opportunities for further research.

Based on the background and review of previous research, this study aims to analyze and compare the performance of the Latent Dirichlet Allocation (LDA) and BERTopic methods in identifying public issues related to the Free Nutritious Food (MBG) program in social media data. This study proposes a structured analysis approach through the stages of data collection, preprocessing, text representation, topic modeling, to coherence score-based evaluation and interpretation of results. The main contribution of this study lies in the comparative application of two different topic modeling methods, namely the probabilistic method and embedding-based methods, in the context of dynamic and unstructured social media data. In addition, this study also provides added value through analysis that not only focuses on the quality of the resulting topics, but also on the ability of each method to represent public issues in a more contextual and relevant manner. Thus, the results of this study are expected to provide methodological recommendations in selecting the appropriate topic modeling model and serve as a reference for further research and policy makers in understanding the dynamics of public opinion towards the MBG program.

Research on topic modeling has grown rapidly in text data analysis, particularly in social media. Embedding-based approaches such as BERTopic have been introduced as methods capable of generating topics with a better level of coherence than traditional methods [1][11]. Several studies have shown that BERTopic has advantages in understanding the semantic context of text data, especially on social media platforms such as Twitter [8]. In addition, the use of contextualized embeddings has also been shown to improve the quality of the generated topics [3].

On the other hand, classical methods such as Latent Dirichlet Allocation (LDA) remain widely used due to their ability to effectively identify latent topic structures [12]. In the evaluation process, the coherence score is used as one of the main metrics to measure topic quality [13]. The development of embedding-based methods is also continuously carried out to improve topic modeling performance [14]. However, the performance of each method is highly dependent on the characteristics of the data used [12][15]. Therefore, a comparative study between LDA and BERTopic is needed in specific contexts, such as public issue analysis in the MBG program.

II. RESEARCH METHODOLOGY

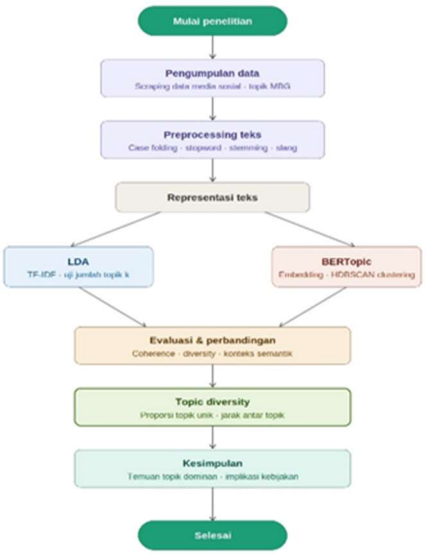

Figure 1. Research Flow

The research began with identifying problems related to analyzing public issues related to the Free Nutritious Food (MBG) program on social media. At this stage, the research objective was determined, namely to compare the performance of the LDA and BERTopic methods in identifying topics.

2.1 Data Collection

In the data collection stage, this study utilized the TikTok platform as the primary source to obtain data related to the Free Nutritious Food (MBG) program issue. Data were collected using scraping techniques based on relevant keywords, resulting in 13,538 data points. The resulting dataset consists of several important attributes, including text (caption or comment content), diggCount (number of likes), replyCommentTotal (number of comment replies), createTimeISO (upload time), uniqueId (user identity), and videoWebUrl and source_file as data sources. In addition, a preprocessing process was carried out that produced derived features such as text_clean, tokens, token_count, tokens_nostop, and text_clean_advanced. The dataset is also equipped with additional information such as text_length, word_count, and temporal attributes (date, month, hour) to support further analysis.

2.2 Text Preprocessing

Text preprocessing is the initial stage in text analysis, which aims to clean and normalize data so that it is ready for further processing. This stage includes case folding, tokenizing, stopword removal, and stemming. In social media data such as TikTok, this process also includes slang normalization to address unstructured language variations. Good preprocessing will improve the quality of text representation and the accuracy of the model used [16]. The collected data is then processed through preprocessing stages to improve data quality. These stages include:

a. Folding case

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

The case folding stage is the process of converting all text to lowercase. This process is carried out as an initial part of text preprocessing to standardize the word formatting in the dataset. Technically, the system will read each character in the document and then convert all capital letters to lowercase using a string transformation function. At this stage, alphabetic characters such as "A" to "Z" are converted to "a" to "z," while non-letter characters can be retained or processed further in the next preprocessing stage. The main purpose of case folding is to avoid differences in word representation due to variations in the use of uppercase and lowercase letters. In text analysis, words such as "Data," "DATA," and "data" actually have the same meaning, but the computer will treat them as different tokens if not normalized. This condition can unnecessarily increase the vocabulary size and cause redundancy in the text feature representation.

b. Tokenizing

The tokenizing stage is the process of breaking down text into smaller units of words or tokens so that the data can be analyzed in a structured manner. At this stage, the system reads each sentence and then separates words based on certain characters such as spaces, punctuation, symbols, or other delimiters. The result of the tokenizing process is a list of tokens that represent each word in the document. For example, the sentence "The MBG program is very helpful for the community" will be converted into tokens such as ["program", "mbg", "very", "helping", "community"]. Technically, tokenizing is a crucial stage in preprocessing because most text mining and natural language processing algorithms cannot directly process raw text in the form of complete sentences. By breaking text into tokens, the system can perform further processes such as filtering, stopword removal, stemming, word weighting using TF-IDF, and even feature formation in topic modeling. In addition, tokenizing also helps the system calculate the frequency of word occurrences and recognize word distribution patterns in documents. Therefore, the quality of the process

c. Stopword removal

```
custom_stopwords = {
    'ya', 'yaa', 'yaaa', 'yg', 'yang', 'di', 'dari', 'dan', 'untuk', 'ini', 'it
    'dengan', 'pada', 'ke', 'adalah', 'juga', 'akan', 'atau', 'tersebut', 'bisa
    'tidak', 'ada', 'sudah', 'saya', 'kita', 'mereka', 'dia', 'kami', 'kalo',
    'kalau', 'nih', 'nah', 'sih', 'dong', 'deh', 'dong', 'lho', 'loh', 'tuh',
    'pak', 'bu', 'bapak', 'ibu', 'mbg', 'bgn', 'ag', 'aja', 'saja', 'banget',
    'banget', 'sangat', 'sudah', 'masih', 'udah', 'akan', 'lagi', 'ttp', 'tetap
    'jd', 'jadi', 'utk', 'utk', 'dgn', 'krn', 'karena', 'jd', 'jdi', 'sdh',
    'bs', 'bisa', 'ga', 'gak', 'nggak', 'nggk', 'gimana', 'gimana', ' Gimana',
    ' gimana', 'dong', 'dongs', 'kak', 'kakak', 'bang', 'teh', 'tuh', 'ni',
    'emang', 'emangnya', 'bener', 'betul', 'sy', 'aku', 'ku', 'mu', 'kau',
    'yaudah', 'ya', 'yaa', 'yok', 'ayo', 'yuk', 'gitu', 'begitu', 'gtu',
    'dm', 'dm', 'gpp', 'gakpapa', 'ok', 'oke', 'sip', 'sipp', 'siapp',
    'please', 'pls', 'tolong', 'tlg', 'thx', 'thanks', 'thank', 'you',
    'wkwk', 'wkwkwk', 'haha', 'hahaha', 'lol', 'lmao', 'btw', 'rt',
    'via', 'amp', 'the', 'and', 'of', 'to', 'is', 'in', 'for', 'on',
    'jd', 'jika', 'bila', 'apabila', 'bahwa', 'sejak', 'sejak', 'antara'
}
```

Figure 2. Stopword Removal

At this stage, a stopword removal process is carried out using a list of custom stopwords compiled according to the characteristics of the social media data. The stopword list includes common Indonesian words such as "dan," "yang," "di," "dari," and "untuk," as well as non-standard and slang words that frequently appear in social media conversations such as "gak," "gak," "aja," "banget," "wkwk," "haha," and "lol." Furthermore, several common English words such as "the," "and," "of," and "is" are also removed because they are deemed not to provide important information to the topic analysis process. Based on the list in the figure, the custom stopwords also include greetings, abbreviations, emotional expressions, and repeated words that frequently appear in social media comments. Technically, the stopword removal process is carried out by comparing each tokenized token with a predetermined list of custom stopwords. If a token is found in the stopword list, it will be removed from the document. Conversely, tokens not included in the list will be retained as important features. The use of custom stopwords aims to make the text cleaning process more relevant to the context of TikTok data, which tends to be informal and unstructured. By eliminating words that do not have important meaning, the text representation becomes more focused on words that represent the main information so that the quality of topic modeling analysis can be improved.

d. Stemming

The stemming stage is carried out to convert each token resulting from the stopword removal process into its basic word form using a stemmer. This process is carried out by iterating on each token, then applying the stemming function to obtain the basic word form so that variations of words with the same meaning can be standardized. If an error occurs in a token during the stemming process, the token is kept unchanged to avoid losing information in the data. The stemming results are then stored as a token list and recombined into complete text for use in the next analysis stage. In addition, a progress indicator is added to monitor the amount of data that has been successfully processed during the process. The stemming stage aims to simplify word variations, reduce redundancy, and increase the consistency of text representation so that the analysis and topic modeling processes can be carried out more effectively.

e. Normalization of slang

In the slang normalization stage, non-standard words, abbreviations, and informal language are replaced with standard word forms using a pre-compiled slang dictionary. The dictionary consists of 423 word entries covering various variations of everyday language commonly used on social media, such as "gk", "gak", and "gak" which are changed to "tidak", and "yg" to "yang". This stage aims to standardize word forms and reduce spelling variations in text data. Technically, the normalization process is carried out by matching each preprocessed token to the slang dictionary. The system will iterate over all tokens, then check whether the token is included in the slang list. If found, the token will be replaced with the standard word form, while tokens not in the dictionary will be retained in their original form. This process is carried out automatically on all documents. Slang normalization is important because social media data generally contains a lot of informal language and abbreviations that can increase noise in the data. By standardizing words that have the same meaning, text representation becomes more consistent, thus improving the quality of analysis and model performance in the next stage[21][22]

f. Bigram/trigram extraction

N-grams are a text processing technique used to capture

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

word sequences within a document. Bigrams and trigrams are able to maintain local context that unigrams cannot. For example, the phrase "nutritious food" will be more meaningful if treated as a single unit. Therefore, the use of n-grams can improve the quality of topic modeling results, especially in short texts such as social media [17]. Bigram and trigram extraction uses the Phrases model to combine words that frequently appear in sequence. The data used comes from normalized tokens, then processed to form two-word (bigram) and three-word (trigram) combinations with a certain minimum occurrence limit. The results show that from 13,574 data, 203 unique bigrams and 75 unique trigrams were formed. Some of the dominant bigrams include "bapak_prabowo", "sehat_selalu", and "terima_kasih", while the trigrams that frequently appear are "di_bawa_pulang" and "terima_kasih_bapak_prabowo". In addition, a visualization is performed to display the top 20 word combinations based on their frequency of occurrence. This process aims to capture a more specific word context so that the analysis becomes more representative.

g. Text Representation

Text representation is the process of converting text data into numerical form so that it can be processed by an algorithm. The LDA method uses the TF-IDF (Term Frequency–Inverse Document Frequency) approach to measure the importance of words in a document. Meanwhile, an embedding-based approach such as BERT is used in BERTopic to capture semantic context more deeply. This model is able to understand the relationship between words in a sentence contextually [15].

1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic-based topic modeling method used to discover latent topics in a collection of documents. LDA assumes that each document consists of a distribution of several topics, while each topic consists of a distribution of words [18]. This method is widely used because it has high interpretability and good computational efficiency, although it still has limitations in capturing semantic context in short texts [19][20]. At this stage, the text data is first converted into a numerical representation using the Term Frequency–Inverse Document Frequency (TF-IDF) method. The TF-IDF method is used to measure the level of importance of a word in a document to the entire collection of documents, where TF indicates the frequency of the word's appearance in the document and IDF is used to reduce the weight of words that appear too often in many documents. The TF-IDF representation produces a numerical matrix that is used as input in the topic modeling process using the LDA method.

$$TF - IDF(t,d) = TF(t,d)x\ \log\left(\frac{N}{df(t)}\right) \quad (1)$$

Information:

- TF(t,d) = word frequency in document
- N = number of documents
- df(t) = number of documents containing the word

2. BERTopic (Embedding)

BERTopic is a modern topic modeling method that combines transformer embeddings, clustering, and c-TF-IDF. This approach enables the formation of more coherent and contextual topics than traditional methods [9][18]. BERTopic is very effective for use on social media data such as TikTok, which is short, dynamic, and unstructured [1], [10][19]. In this approach, each document is first converted into a numeric vector using sentence embedding, then automatically grouped into topics using a density-based clustering algorithm such as HDBSCAN, without the need to specify the number of topics in advance. Mathematically, the embedding representation process can be expressed as:

$$V_d = f(d) \quad (2)$$

Information:

- Vd = document embedding vector d
- f(d) = embedding function (e.g. transformer-based model like BERT)

Next, the clustering process is carried out by grouping documents based on the proximity of the distance in the vector space, which is generally calculated using cosine similarity:

$$sim(x,y) = \frac{x.y}{||x|||y||} \quad (3)$$

Once the clusters are formed, each topic is represented using class-based TF-IDF (c-TF-IDF) to extract the most representative words in each topic:

$$\text{c\_TFIDF}_{t,c} = \frac{f_{t,c}}{\sum f_{t,c}} x \log\left(\frac{N}{df_t}\right) \quad (4)$$

Information:

- ft,c= frequency of word t in cluster c
- N = total number of documents
- dft = number of clusters containing word t

h. Evaluation and Comparison

Evaluation is done to measure the quality of the topic using several metrics:

1. Coherence Score

$$C = \frac{1}{M}\sum log\left(\frac{D(w_i,w_j)+1}{D(W_j}\right) \quad (4)$$

2. Topic Diversity

The topics generated by the topic modeling model. This metric aims to ensure that each topic has a unique word representation and does not experience excessive word overlap. The higher the topic diversity value, the more diverse the words used in each topic, thus improving the quality of topic separation. Conversely, a low topic diversity value indicates that many words appear repeatedly across several topics, so the resulting topics tend to be less specific and similar to each other. Mathematically, topic diversity is calculated based on the proportion of unique words to the total words from all top topics generated by the model. In the formula, represents the set of words in the i-th topic, k indicates the number of topics, and n is the number of top words retrieved from each topic. The union operator is used to calculate the total unique words from all topics. This value is then divided by the product of the number of topics and the number of top words in each topic (k × n). Thus, topic diversity can be

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

used to evaluate the extent to which the model is able to generate diverse, representative topics with minimal overlap between topics. In general, topic diversity is calculated based on the proportion of unique words to the total words appearing in a number of top topics, which can be formulated as:$W_i \cup_1^k = 1W_i$

$$TD = \frac{|U_I^K = W_i|}{k\ x\ n} \quad (5)$$

Information:
- TD = topic diversity
- Wi = set of words on topic i
- k = number of topics
- n = number of top words on each topic

A high topic diversity value indicates that the generated topics have more diverse words, thus indicating better topic modeling quality.

III. Results and Discussion

3.1 LDA(TF-IDF)

In the topic modeling stage using the Latent Dirichlet Allocation method, the text data is first converted into a Document-Term Matrix using CountVectorizer with parameters max_df=0.95, min_df=2, and max_features=10000. The transformation results in a matrix of size (6787, 4599), which shows that there are 6,787 documents and 4,599 unique word features used in the topic modeling process.

```
Training LDA dengan k=5...
  Perplexity: 764.99
  Coherence (C_v): 0.5068
  Log-Likelihood: -606265.69
  Topic Diversity: 0.9000
  Top 5 kata topik 0: gizi, pangan, pelayanan, mokel, tahu

Training LDA dengan k=8...
  Perplexity: 799.22
  Coherence (C_v): 0.5098
  Log-Likelihood: -610262.79
  Topic Diversity: 0.9000
  Top 5 kata topik 0: tahu, min, bpk_prabowo, omon, masya_allah

Training LDA dengan k=10...
  Perplexity: 828.40
  Coherence (C_v): 0.4782
  Log-Likelihood: -613536.58
  Topic Diversity: 0.9100
  Top 5 kata topik 0: tahu, pemimpin, wo, kapan, jangan
```

Figure 3. Coherence Score Matrix

Next, LDA model training was conducted with varying topic counts of 5, 8, 10, 12, 15, and 20 topics to find the optimal model. Evaluation was performed using perplexity, coherence score (C_v), log-likelihood, and topic diversity metrics. The test results showed that the model with 8 topics produced the highest coherence score of 0.5098 with a topic diversity of 0.9000, so it was selected as the best model because it was able to produce more coherent and representative topics compared to other topic counts.

	k	perplexity	coherence	log_likelihood	topic_diversity
0	5	764.987803	0.506777	-606265.686932	0.900000
1	8	799.220142	0.509753	-610262.791483	0.900000
2	10	828.395877	0.478227	-613536.578687	0.910000
3	12	851.365677	0.465593	-616033.878700	0.933333
4	15	876.960178	0.475153	-618738.377923	0.886667
5	20	909.348790	0.430043	-622049.821984	0.955000

Table 1. LDA Model Evaluation Results Based on Variations in the Number of Topics

The evaluation results of the Latent Dirichlet Allocation model were then stored in a dataframe to facilitate analysis and comparison between models. Based on the evaluation results, the model with 8 topics produced the highest coherence score of 0.5098 with a topic diversity of 0.9000. Meanwhile, the model with 5 topics had the lowest perplexity score of 764.99. However, as the number of topics increased to 20, the perplexity value tended to increase and the coherence score decreased. This indicates that too many topics can reduce the quality of semantic relationships between words in the topic. Therefore, the model with 8 topics was chosen as the optimal model because it provides the best balance between topic quality and data representation.

3.2 Topic (Embedding)

In the topic modeling stage using the BERTopic method, this study uses the multilingual SentenceTransformers paraphrase-multilingual-MiniLM-L12-v2 embedding model to represent text into semantic vectors. The modeling process is carried out with the parameters language='indonesian', nr_topics='auto', top_n_words=15, and calculate_probabilities=True. The stages start from document transformation into embedding, dimensionality reduction using UMAP, density-based clustering, to topic extraction using c-TF-IDF. The modeling results show that BERTopic successfully reduced 101 initial topics to 25 topics, with 24 main topics after removing outliers, and produced 2,345 outlier documents. These results indicate that BERTopic is able to identify public issues related to the MBG program more contextually and representatively in short and unstructured social media data.

```
BERTopic Coherence (C_v): 0.4133
BERTopic Topic Diversity: 0.7667
```

Figure 4. BERTopic Coherence and Topic Diversity Values

Based on the evaluation results, the BERTopic method produced a coherence score (C_v) of 0.4133 and a topic diversity score of 0.7667. The coherence score indicates that the resulting topics have a fairly good semantic relationship, while the topic diversity score indicates a relatively high diversity of words between topics. These results indicate that BERTopic is capable of producing fairly representative and contextual topics in identifying public issues related to the MBG program in social media data.

3.3 Model Evaluation

	Metrik	LDA	BERTopic
0	Coherence Score (C_v)	0.5098	0.4133
1	Perplexity	799.22	N/A
2	Topic Diversity	0.9000	0.7667
3	Silhouette Score	N/A	-0.0510
4	Jumlah Topik	8	24
5	Outlier (%)	0.0%	34.6%
6	Jumlah Dokumen	6787	6787

Table 2. Model Evaluation

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

Based on the evaluation results, the LDA method produced a coherence score of 0.5098, higher than BERTopic's 0.4133, indicating that the topics in LDA have better word relationships. In the topic diversity metric, LDA also obtained a higher score of 0.9000 compared to BERTopic's 0.7667, thus indicating better word diversity between topics. Meanwhile, BERTopic produced 24 topics with outliers of 34.6%, while LDA produced 8 topics without outliers. In addition, BERTopic obtained a silhouette score of -0.0510, indicating that the quality of cluster separation is still less than optimal. Overall, LDA showed more stable performance based on the evaluation metrics, while BERTopic was superior in generating contextual topics in social media data.

3.4 Topic Diversity

Topic diversity was implemented by calculating the number of unique words across all topics and then dividing them by the total number of words used. Furthermore, a similarity analysis was performed between topics to determine the extent to which the words appearing in each topic overlap. A higher topic diversity value indicates that the resulting topics have more diverse and representative words, thereby reducing overlap between topics. Based on the evaluation results, the LDA method produced a higher topic diversity value than BERTopic. In the LDA model, the formula topic diversity value increased to around 0.92 at k = 12 and tended to be stable at subsequent topic counts, while the pairwise topic diversity value approached 1.0, indicating a very good level of word difference between topics. Meanwhile, BERTopic produced a topic diversity value of 0.7667, indicating that the word diversity between topics was still below LDA. These results indicate that LDA is more capable of producing diverse topics with minimal overlap, while BERTopic is superior in producing contextual topics in social media data.

	Model	Jumlah Topik	TD Formula	TD Pairwise	Unique Words
0	LDA (k=5)	5	0.760000	0.895533	38
1	LDA (k=8)	8	0.875000	0.976817	70
2	LDA (k=10)	10	0.890000	0.983863	89
3	LDA (k=12)	12	0.925000	0.990076	111
4	LDA (k=15)	15	0.906667	0.988577	136
5	LDA (k=20)	20	0.890000	0.988681	178
6	BERTopic (80 topics)	80	0.823750	0.995835	659

Table 3. Comparison of Topic Diversity

Based on the results of the topic diversity comparison, the LDA and BERTopic methods show different characteristics in producing topic diversity. In the LDA model, the TD Formula value increases with the increase in the number of topics, starting from 0.7600 at k = 5 to reach the highest value of 0.9250 at k = 12 with 111 unique words. After that, the diversity value tends to decrease slightly at k = 15 and k = 20. Meanwhile, the Pairwise TD value in LDA also increases from 0.8955 to close to 1.0, indicating that the topics have a very good level of word differentiation and minimal overlap. On the other hand, BERTopic with 80 topics produces a TD Formula value of 0.8238 and a TD Pairwise value of 0.9958 with a total of 659 unique words. These results indicate that BERTopic is able to produce a much larger number of unique words and has a very high difference between topics, but the TD Formula value is still below the optimal LDA at k = 12. Overall, LDA produces a more stable and balanced topic diversity, while BERTopic produces more and highly semantically diverse topics in social media data.

IV. CONCLUSION

Based on the research results, it can be concluded that the Latent Dirichlet Allocation (LDA) and BERTopic methods have distinct characteristics and advantages in identifying public issues related to the Free Nutritious Food (MBG) program in TikTok social media data. The analysis process was carried out through data collection, text preprocessing, text representation, topic modeling, and evaluation using coherence scores, topic diversity, and manual interpretation.

1. The results showed that the LDA method produced better evaluation performance than BERTopic based on coherence score and topic diversity. The optimal LDA model with 8 topics obtained a coherence score of 0.5098 and a topic diversity of 0.9000, indicating that the resulting topics had good word associations and high topic diversity. Furthermore, LDA was able to produce more stable topics with minimal overlap between topics.
2. On the other hand, BERTopic demonstrated superiority in generating more contextual and semantic topics due to its use of transformer-based embedding. BERTopic successfully identified 24 key topics from short, unstructured social media data. Although the coherence score and topic diversity obtained were lower than LDA, at 0.4133 and 0.7667, respectively, this method was able to capture the context of public issues more deeply and representatively.
3. Overall, this study shows that LDA is superior in terms of model stability, word association quality, and topic diversity, while BERTopic is more effective in understanding semantic context in social media data. Therefore, the choice of topic modeling method needs to be tailored to the analysis objectives and the characteristics of the data used. The results of this study are expected to serve as a reference in the development of social media-based public issue analysis and provide methodological recommendations for further research in the fields of topic modeling and natural language processing.

Future research is recommended to use a larger dataset from various social media platforms to provide a more representative public issue analysis. Furthermore, this research can be developed by comparing it with other topic modeling methods, such as Top2Vec or NMF, to achieve more optimal results. Preprocessing that is more adaptive to social media slang and language also needs to be improved. Future research could also incorporate sentiment analysis to gain a more in-depth and comprehensive understanding of public opinion trends toward the MBG program.

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

REFERENCES

[1] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2020.

[2] P. Koochemeshkian and N. Bouguila, "Integration of Neural Embeddings and Probabilistic Models in Topic Modeling Integration of Neural Embeddings and Probabilistic Models in Topic Modeling," Appl. Artif. Intel., vol. 38, no. 1, 2024, doi: 10.1080/08839514.2024.2403904.

[3] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a Hot Topic : Contextualized Document Embeddings Improve Topic Coherence," pp. 759–766, 2021.

[4] AB Dieng and DM Blei, "Topic Modeling in Embedding Spaces," 2017.

[5] S. Koltcov, A. Surkov, V. Filippov, and V. Ignatenko, "Topic models with elements of neural networks : investigation of stability, coherence, and determining the optimal number of topics," pp. 1–41, 2024, doi: 10.7717/peerj-cs.1758.

[6] L. Yijia, "Comparison of LDA and BERTopic in News Topic Modeling : A Case Study of The New York Times' Reports on China," vol. 7, no. March, pp. 47–51, 2024, doi: 10.55014/pij.v7i3.616.

[7] MD Pratiwi, KD Tania, S. Informasi, and U. Sriwijaya, "Knowledge Discovery Through Topic Modeling on GoPartner User Reviews Using BERTopic, LDA, and NMF," vol. 9, no. 1, pp. 1–7, 2025.

[8] Egger, J. Yu, and J. Yu, "A Topic Modeling Comparison Between LDA , NMF , Top2Vec , and BERTopic to Demystify Twitter Posts Making Sense of Social Media Using," vol. 7, no. May, pp. 1–16, 2022, doi: 10.3389/fsoc.2022.886498

[9] Luiz and M. Owa, "Identification of Topics from Scientific Papers through Topic Modeling," pp. 541–548, 2021, doi: 10.4236/ojapps.2021.114038.

[10] CB Pavithra and J. Savitha, "Topic Modeling for Evolving Textual Data Using LDA , HDP , NMF , BERTOPIC , and DTM With a Focus on Research Papers," vol. 5, no. 2, 2024, doi: 10.37802/joti.v5i2.618.

[11] P. Resnik, "Improving Neural Topic Models using Knowledge Distillation," pp. 1752–1771, 2020.

[12] A. Farea, S. Tripathi, G. Glazko, and F. Emmert-streib, "Engineering Applications of Artificial Intelligence Investigating the optimal number of topics by advanced text-mining techniques : Sustainable energy research," Eng. Appl. Artif. Intel., vol. 136, no. PA, p. 108877, 2024, doi: 10.1016/j.engappai.2024.108877.

[13] X. Wu, T. Nguyen, and AT Luu, "and challenges," Artif. Intel. Rev., vol. 57, no. 2, pp. 1–30, 2024, doi: 10.1007/s10462-023-10661-7.

[14] MR Haas and A. Heijn, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," pp. 1–3, 2020.

[15] MC Kenton, L. Kristina, and J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. MLM, 1953.

[16] E. Taheri and JL Junkins, "How Many Impulses Redux".

[17] M. Röder and A. Hinneburg, "Exploring the Space of Topic Coherence Measures".

[18] P. Aprilio, PS Nugraha, and H. Fahmi, "Hybrid Feature Combination of TF-IDF and BERT for Enhanced Information Retrieval Accuracy," vol. 08, no. 01, pp. 8–15, 2025.

[19] DM Blei, AY Ng, and MI Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.

[20] I. Computer and T. Hofmann, "Probabilistic Latent Semantic Indexing," pp. 50–57.

[21] COO Optimization, "Topic Modeling as Multi-Objective Contrastive Optimization," pp. 1–20, 2024.

[22] 21] PMSArdinata, AAJ Permana, and INSW Wijaya, "IDENTIFICATION AND NORMALIZATION OF SLANG TEXT WITH," vol. 21, no. 1, 2024.

[23] [22] RLNurdiansyah and KE Dewi, "KOMPUTA: Scientific Journal of Computers and

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN:2776-3234

Informatics THE EFFECT OF INFORMATION GAIN AND WORD NORMALIZATION ON ASPECT-BASED SENTIMENT ANALYSIS KOMPUTA: Scientific Journal of Computers and Informatics," vol. 12, no. 2, 2023.

[24] R. Li, F. González-pizarro, L. Xing, G. Murray, and G. Carenini, "Diversity-Aware Coherence Loss for Improving Neural Topic Models," vol. 2, pp. 1710–1722, 2023.

[25] R. Li, F. González-pizarro, L. Xing, G. Murray, and G. Carenini, "Diversity-Aware Coherence Loss for Improving Neural Topic Models," vol. 2, pp. 1710–1722, 2023.

JISA (Journal of Informatics and Science) (e-ISSN: 2614-8404) is published by the Informatics Engineering Study Program, Trilogi University under Creative Commons Attribution-ShareAlike 4.0 International License.