Journal of Information Systems and Informatics
Vol. 7, No. 3, September 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i3.1266 Published By DRPM-UBD

Reducing Semantic Distortion of Multiword Expressions for Topic Modeling with Latent Dirichlet Allocation

Widya Astuti Sitopu[1], Erna Budhiarti Nababan[2], Mohammad Andri Budiman[3]

[1,2,3] Master of Data Science and Artificial Intelligence, Faculty of Computer Science and Technology, Universitas Sumatera Utara, Medan, Indonesia
Email: [1]widyaastuti@students.usu.ac.id, [2]ernabrn@usu.ac.id, [3]mandrib@usu.ac.id

Abstract

The Makan Bergizi Gratis (MBG) is one of the Indonesian government's priority initiatives that has received significant coverage in online media. To understand the main themes within these narratives, this study applies topic modeling using Latent Dirichlet Allocation (LDA). However, the results of topic modeling are highly influenced by the preprocessing stage, particularly in handling multiword expressions (MWEs) such as named entities, collocations, and compound words. This study compares two preprocessing approaches: basic and extended, with the latter involving the masking of MWEs. Experimental results show that the extended preprocessing model achieved the highest coherence score of 0.5149 at K=22K = 22K=22, with four other scores also exceeding 0.496, whereas the basic preprocessing model only reached a maximum of 0.3932 at K=10K = 10K=10. Furthermore, cosine similarity scores between topics in the extended model were lower (maximum 0.7406) than in the basic model (maximum 0.8244), indicating that the topics produced were more diverse and less overlapping. These findings highlight the importance of preprocessing strategies that preserve phrase-level meaning to reduce semantic distortion and improve topic coherence and representation-particularly in analyzing media discourse on public policy programs such as MBG.

Keywords: Multiword Expression (MWE), Text Preprocessing, Topic Modeling, Latent Dirichlet Allocation (LDA), Topic Coherence

1. INTRODUCTION

The Free Nutritious Meal Program (MBG) is one of the government's priority programs that has become a national discourse and has been widely discussed in recent times [1]. According to the Ministry of Finance (2024), this program aims to improve the quality of human resources by strengthening nutrition for school children [2]. Along with its implementation, the MBG has been in the media spotlight because it is not free from various challenges in the field. News about free nutritious meals has pros and cons, ranging from support for the benefits of this program for school children, to criticism of its implementation which is considered less than optimal [3], [4].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Online news, as a primary medium for shaping public opinion, presents information about the Free Meal Program (MBG) quickly, broadly, and regularly updated [5]. In recent months, various national and local media outlets have published numerous articles discussing the program's policies, implementation, and impact. The high volume and variety of coverage demonstrates that the MBG issue touches on broader social, economic, and political dimensions. Understanding how the media conveys this issue, particularly the topics covered, is crucial for assessing public opinion on the free nutritious meal program [6].

To gain a deeper understanding of public opinion, analysis of MBG news can focus on topic modeling. Topic modeling is a text mining method capable of identifying and analyzing hidden topics within a collection of text documents (in this case, news related to MBG) to aid in understanding and managing large-scale data [7]. This method allows for a more structured understanding of the main themes discussed in a collection of text without having to read the entire document directly. In the context of news about MBG, topic modeling can reveal how the media constructs narratives, the dominant topics that frequently arise, and the dynamics of the issue's development over time [8].

Text analysis representation relies heavily on data preprocessing. The rich morphological structure of Indonesian, particularly word formation involving numerous affixations and reduplications, presents challenges that require special attention during the text preprocessing stage. One important initial process in this stage is tokenization, which separates text into word units [9]. Tokenization does not consider the context of the phrase, thus separating important word combinations such as "eat nutritiously free" into individual word units that lose their original semantic meaning (are distorted). In addition to tokenization, the stemming process also plays a significant role in shaping text representation. Overly aggressive or inaccurate stemming can truncate words inappropriately and cause changes in meaning. For example, the word "*pelaksanaan*" stemmed as "*laksana*" can cause confusion in topic interpretation. This problem is exacerbated by the inability to recognize multiword expressions (MWE), which are word combinations whose meaning must be processed as a single unit. If a phrase such as "eat nutritiously free" is separated and processed word by word, the intended meaning of the policy is lost and changed [10]. Additionally, there are also specific names or entities (types of MWE) in the text that need to be protected from distortion during the preprocessing stage. As explained by Amalia et al. (2020), efforts to maintain semantic representation, including handling multiword expressions, are crucial, especially in small corpora and specific domains. Mishandling these elements can result in the loss of important information or changes in meaning, which will affect topic modeling results [11].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Although research on topic modeling has developed extensively, many studies tend to use deep learning-based approaches such as transformers (BERTopic), which internally can understand the context more deeply without focusing on advanced preprocessing. However, according to research by Bonetti et al. (2023), the analysis results of traditional and semi-traditional approaches such as Latent Dirichlet Allocation (LDA) are highly dependent on strong data preprocessing. According to Kresnawan et al. (2021), LDA is used to extract the main topic from a collection of documents by modeling each document as a probability distribution over topics, and each topic as a distribution of words. This process produces a number of topics, each consisting of words and their probability values [12]. However, although very comprehensive, this study does not specifically discuss the handling of multiword expressions (MWE) or entity protection using named entity recognition (NER) in preprocessing [13]. Then, research by Khairova et al. (2024) used LDA with Collapsed Gibbs Sampling to analyze news related to the Ukraine war, acknowledging the importance of preprocessing multi-word combinations. However, MWE handling was only mentioned without implementation and was not accompanied by entity protection or a deep stemming approach. Meanwhile, Cheevaprawatdomrong et al. (2021) emphasized the importance of phrase meaning, as their research showed that MWE handling can improve LDA performance, both in terms of topic coherence and semantic representation [14].

Based on the previous description, this study aims to compare the results of multiword expression handling to reduce meaning distortion in topic modeling for news about free nutritious meals. This study also aims to demonstrate that implementing appropriate preprocessing steps can produce more meaningful analysis and emphasize the importance of a data-centric approach in text mining to support model performance [15].

2. METHODS

2.1. Data Collection

The data in this study were obtained through a web scraping process on Indonesian-language online news sites. The online news media used were Kompas (https://www.kompas.com/), Tempo (https://www.tempo.co/), and Detik (https://www.detik.com/) with the tag "makan-bergizi-gratis". Data collection was carried out for the period May 23, 2024 to May 23, 2025. The tag "makan-bergizi-gratis" produced 2003 news documents from Kompas media, 1332 news documents from Tempo media, and 1778 news documents from Detik media, which were combined into 5113 news documents. The following is an example of the collected data. The following shows the general architecture of this research.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 1. General Architecture

In this study, the data collection process was conducted using web scraping techniques. Scraping was carried out in two stages: extracting news links and extracting news content [16]. The following displays the data from the collected documents.

Table 2. News Document Content

Index	Title	Content (Indonesia)	Media
0	Dua Fraksi DPR Puja-puji Program Makan Bergizi Gratis Prabowo	*Fraksi Golkar mengapresiasi pemerintah karena visi misi dan program unggulan presiden dan wakil presiden terpilih Prabowo Subianto-Gibran Rakabuming Raka, dimasukkan dalam Kerangka Ekonomi Makro dan Pokok-pokok Kebijakan Fiskal (KEM-PPKF) RAPBN 2025. Salah satunya program makan bergizi gratis untuk anak sekolah. Anggota Komisi IX DPR Dewi Asmara mewakili Fraksi Golkar mengatakan hal ini*	detik

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

		penting dilakukan agar pemerintahan baru bisa langsung melaksanakan programnya dengan dukungan anggaran yang…	
…	…	…	…
5112	Peringatan Wakil Bupati Karanganyar untuk ASN di Bumi Intanpari...	*KARANGANYAR, KOMPAS- Wakil BupatiKaranganyar, Adhe Eliana, mengingatkan para Aparatur Sipil Negara (ASN) untuk bersiap menghadapiefisiensi anggaran. Peringatan tersebut disampaikan saat memimpin apel perdana di lingkungan Pemerintah Kabupaten (Pemkab) Karanganyar pada Senin (24/2/2025). "Kalau efisiensi ini sudah keluar, pastinya semuanya akan menjadi polemik dan itu pasti. Dan kita pasti siap tidak siap harus siap," ujar Adhe dalam sambutannya. Baca juga: Analisis dan Dampak Kebijakan Penghematan Anggaran Prabowo Subianto... Adhe juga menekankan pentingnya kesabaran dan semangat ASN….*	kompas

2.4. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) process was conducted to understand the dataset's structure, quality, and distribution before proceeding to preprocessing and modeling. The main steps are as follows:

1) Check Data Structure and Type

The initial step involved examining the dataset's structure and attribute types to ensure that each variable had the correct data format and to identify potential inconsistencies.

```
 0   judul   5093 non-null   object
 1   isi     4999 non-null   object
 2   media   5113 non-null   object
dtypes: object(3)
memory usage: 120.0+ KB
```

Figure 2. Data Structure and Types

2) Check for Duplicate Values.

A duplication check was performed to identify and remove redundant entries. The results revealed that two records contained identical information, indicating the need for duplicate handling during the preprocessing stage. The findings are summarized in Table 3[17].

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Table 3. Duplicate Value Check Results

Tndex	Title	Content (Indonesia)	Media
2385	Demonstrasi BEM SI Indonesia Gelap, Mahasiswa Bersiap ke Patung Kuda Monas	*Baca berita dengan sedikit iklan,klik di sini TEMPO.CO,Jakarta- Hingga Senin siang pukul 13.55 WIB, demonstrasi bertajukIndonesia Gelapdi kawasan Patung Kuda Monas, Jakarta Pusat, belum juga berlangsung. Tampak puluhan mahasiswa Universitas Pembangunan Nasional Veteran Jakarta dan Universitas Muhammadiyah Prof. Dr. Hamka (Uhamka) berteduh di area parkir Masjid Rihlatul Jannah dari guyuran air hujan sambil menunggu unjuk rasa dimulai. Baca berita dengan sedikit iklan, klik di sini*……	tempo
2386	Demonstrasi BEM SI Indonesia Gelap, Mahasiswa Bersiap ke Patung Kuda Monas	*Baca berita dengan sedikit iklan,klik di sini TEMPO.CO,Jakarta- Hingga Senin siang pukul 13.55 WIB, demonstrasi bertajukIndonesia Gelapdi kawasan Patung Kuda Monas, Jakarta Pusat, belum juga berlangsung. Tampak puluhan mahasiswa Universitas Pembangunan Nasional Veteran Jakarta dan Universitas Muhammadiyah Prof. Dr. Hamka (Uhamka) berteduh di area parkir Masjid Rihlatul Jannah dari guyuran air hujan sambil menunggu unjuk rasa dimulai. Baca berita dengan sedikit iklan, klik di sini*……	tempo

3) Check Text Length Statistics

Descriptive statistics were calculated to analyze the distribution of text lengths within the dataset. This analysis provides insights into the variability of document sizes and guides appropriate text segmentation or tokenization strategies. The results are illustrated in Figure 3 [18].

4) Check for Unusual Characters

This step involved identifying and counting unusual or non-standard characters, including punctuation marks, numeric symbols, emojis, non-alphanumeric, and

non-Latin characters. The findings support the design of a targeted text-cleaning procedure during the preprocessing stage [19].

```
count     5113.000000
mean      2478.860160
std       1197.170969
min          1.000000
25%       1895.000000
50%       2382.000000
75%       2939.000000
max      14271.000000
Name: panjang_teks, dtype: float64
```

Figure 3. Results of Text Length Statistics Check

5) Check the Potential for Multiword Expressions (MWE).

To investigate potential multiword expressions, word cloud visualizations were generated for the most frequent bigrams and trigrams. This helped identify compound expressions and contextual word relationships that could enrich the feature representation. The visualization results are shown in Figure 4 [20].

Figure 4. MWE Potential Check Results

2.5. Data Preprocessing

Basic Text Preprocessing consists of eight stages: handling missing values, handling duplicate values, text cleaning, lowercasing, tokenization, stopword removal, stemming, and feature extraction. Meanwhile, one important stage is added before the final tokenization process in Extended Text Preprocessing:

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

multiword expression detection with IndoBERT. The final result of this process is a corpus, which is a representation of each document in the form of a list of pairs (token, count) that show how often each token appears in the document. In the next stage, these tokens will be mapped to token_id through a dictionary, so that the corpus can be used as numeric input for the Latent Dirichlet Allocation (LDA) model. The following table shows an example of the corpus results generated from basic text preprocessing and extended text preprocessing. The results of extended text preprocessing show the presence of multiword expression (MWE) tokens, such as "makan gizi gratis" and "sultan b najamuddin," which were identified through the masking and phrase detection stages. The presence of MWE is expected to improve the topic coherence generated by the model, as shown in Table 4.

Table 4. Corpus Contents of Basic and Extended Text Preprocessing

No	corpus basic text preprocessing	corpus extended text preprocessing
347	[('gizi', 2), ('gratis', 1), ('makan', 1), ('penuh', 1), ('program', 1), ('salah', 1), ('satu', 2), ('menu', 1), ('siap', 1), ('mbg', 2), ('layan', 1), ('dapur', 2), ('depok', 2), ('sppg', 1), ('bayun', 1), ('depok-', 1), ('mpg', 1), ('tapos', 2)]	[('program', 1), ('salah', 1), ('satu', 2), ('menu', 1), ('ada', 1), ('lihat', 1), ('siap', 1), ('mbg', 2), ('layan', 1), ('dapur', 2), ('makan gizi gratis', 1), ('depok', 1), ('sppg', 1), ('penuh gizi', 1), ('bayun', 1), ('mpg', 1), ('tapos', 1), ('tapos depok depok-', 1)]

The final corpus results are shown in Table 5 is the form of a Bag of Words (BoW). At this stage, each text document that has undergone text preprocessing is converted into a list of (token_id, count) pairs, where the token_id is obtained from a dictionary that maps each unique token to a numeric ID.

Table 5. Bag of Words Results from Basic and Extended Text Preprocessing

No	corpus basic text preprocessing	corpus extended text preprocessing
347	[(0,2),(1,1),(3,1),(47,1),(5,1),(22,1),(23,2),(48,1),(49,1),(30,2),(50,1),(51,2),(52,2),(53,1),(54,1),(55,1),(56,1),(57,2)]	[(5,1),(22,1),(23,2),(48,1),(58,1),(59,1),(49,1),(30,2),(50,1),(51,2),(17,1),(52,1),(53,1),(60,1),(54,1),(56,1),(57,1),(61,1)]

The count of (token_id, count) indicates the frequency of occurrence of the token in the document. The results of extended text preprocessing show the presence of multiword expression (MWE) tokens such as "makan gizi gratis" or "sultan b najamuddin" that have been combined and given their own token_id. This numeric representation is then used as input for topic modeling with Latent Dirichlet Allocation (LDA).

To determine the optimal number of topics, experiments were conducted by trying various values of the number of topics (K), ranging from 2 to 64. At each iteration, an LDA model was built with parameters passes=20 and iterations=400 to ensure a sufficiently in-depth training process, and alpha and eta were set to 'auto' to allow the model to adjust its topic distribution based on priors (learn the topic distribution itself). The model also used the option per_word_topics=True so that it could produce a topic distribution per word. The final result of LDA modeling is two important components: the topic distribution per document and the word distribution per topic. This information is then used to compare topic modeling results between basic text preprocessing and extended text preprocessing, specifically to analyze the effect of multiword expression masking on the coherence of the resulting topics.

2.7. Model Evaluation

In this study, the evaluation was conducted using two main approaches: topic coherence and cosine similarity between topics [9]. These two metrics were used to assess the extent to which the LDA model was able to generate meaningful, non-overlapping topics [21].

3. RESULTS AND DISCUSSION

3.1. Determining the Optimal K (Number of Topics) Value

Determining the number of topics (K) is a crucial step in topic modeling using Latent Dirichlet Allocation (LDA). Too small a K value can result in overly general topics, while too large a K value can result in overlapping or meaningless topics. Therefore, an experiment was conducted to evaluate the coherence score of the LDA model with various numbers of topics [10]. The experiment was conducted by building an LDA model for K=2–64 and calculating the coherence score (c_v) for each K value, then visualizing it in a graph. Figure 5 shows the coherence results with a range of K values from 2 to 63 from basic text preprocessing and extended text preprocessing.

Figure 5 show that in the basic text preprocessing approach, the coherence value increases with the number of topics, reaching a peak at K=10 with a coherence value of 0.3932. After that, the coherence value begins to fluctuate and shows no significant improvement. Meanwhile, in the extended text preprocessing approach (with multiword expression masking), the coherence value is already relatively high at the start, ranging from 0.44–0.49, and increases significantly in the K=20–23 range, peaking at K=22, reaching a maximum value of 0.5149. Figure 5 shows that the cleaned dataset from extended text preprocessing consistently produces higher

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

coherence than basic text preprocessing at almost all K values, especially around the optimal number of topics.

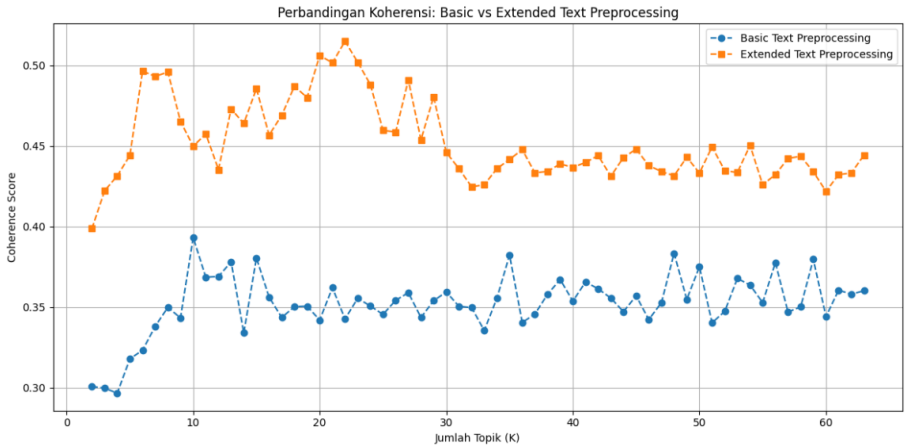

Figure 5. Coherence Results from Data to Find Optimal K

3.2. Topic Distribution per Word

After determining the optimal number of topics (10 for the model with basic text preprocessing and 22 for the extended text preprocessing), the next step was to analyze the distribution of key words representing each topic. The following shows the distribution of topics by word, Table 6 show the top ten words from each topic for the LDA model with basic text preprocessing.

Table 6. Top Ten Words from the Basic Text Preprocessing Model Topics

Topic	Key Words (Indonesia)
1	*indonesia, papua, program, perintah, ajar, tolak, tni, gratis, didik, masyarakat*
2	*mbg, siswa, makan, racun, sekolah, gizi, menu, sppg, dapur*
3	*gizi, program, makan, mbg, gratis, dad, bgn, badan, nasional, laksana*
4	*kantin, surabaya, dagang, sumenep, henti, mbg, makan, sekolah, gratis, jual*
…	…
10	*rp, gizi, program, anggar, mbg, makan, triliun, juta, gratis*

The distribution of these words indicates that the topics generated by the model are closely related to the issue of the free nutritious meal program (MBG). This is evident from the dominance of words such as "eat," "nutrition," "free," "program," and "school" that recur across various topics. Furthermore, there are topics that highlight specific entities or regions, such as "Papua," "Jakarta," or figures like "Prabowo" and "Gibran," indicating a geographic and political aspect in the analyzed documents.

The occurrence of words such as "sppg," "kitchen," and "kantin" also indicates a focus on the technical implementation of the nutritious meal program. Thus, the topics obtained from the LDA model are generally consistent and relevant to the theme raised in the document corpus, namely the MBG program, which has become a public discourse. Table 7 shows the distribution of topics by word, showing the top ten words for each topic for the LDA model with extended text preprocessing.

Table 7. Top Ten Words from Topics in the Extended Text Preprocessing Model

Topic	Key Words (Indonesia)
1	*makan, program, gizi, umkm, gratis, mbg, usaha, jadi, koperasi, libat*
2	*desa, tni, pangan, program, gizi, tani, bangun, siap, bahan, jadi*
3	*manggarai, sikka, domba, khoirudin, bangka, ganjar, karya, persero, kereta, labu*
4	*prabowo, menteri, presiden, perintah, indonesia, negara, tahun, efisiensi, kerja, kabinet*
5	*makan, gizi, anak, gratis, jadi, program, sehat, menu, protein, beri*
…	…
22	*jakarta, makan, coba, uji, gizi, gratis, per, porsi, heru*

These topics show that the LDA model is able to recognize important themes in the corpus, such as free nutritious meal programs (Topics 1, 5, 6, 8), political and government issues (Topics 4, 13, 14, 15), up to certain geographic areas (Topics 3, 10) and child health and nutrition (Topics 5, 7, 11, 20). This indicates that the extended text preprocessing process successfully lifts relevant MWE (entities and word combinations) to support more semantic and informative topic modeling.

3.3. Distribusi Topic in Document

After the LDA model is built with the optimal number of topics, an analysis is performed to see how the topic distribution appears in each document [22]. This process is done by taking a representation of the topic distribution from each document. The output of the model is a probability distribution for each topic that reflects the extent to which a document is affiliated with certain topics [23]. Figure 6 shows the topic distribution across all documents from the cleaned dataset with basic text preprocessing.

The visualization results show that Topic 7, Topic 2, and Topic 5 are the most dominant topics in the corpus, with average probabilities of approximately 0.16, 0.16, and 0.15, respectively. This indicates that these three topics appear consistently and significantly across most documents, making them key representatives of the themes contained in the data. Meanwhile, Topic 9, Topic 1, and Topic 6 also have significant contributions, but slightly lower than the three

main topics. Conversely, Topic 3, Topic 4, and Topic 0 have the lowest average probabilities, each below 0.06, with Topic 3 appearing the least frequently (around 0.02). This suggests that these topics likely appear in only a small proportion of documents, or only serve as supporting context for the main theme.

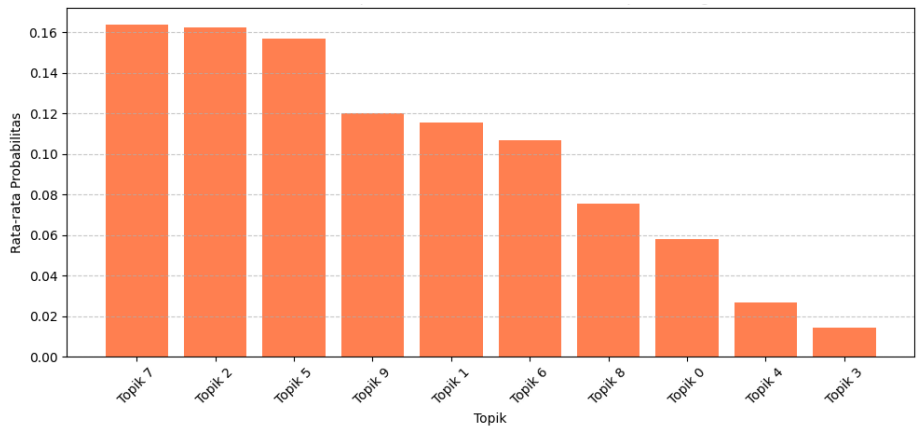

Figure 6. Topic Distribution Across Basic Text Preprocessing Documents

Figure 7 illustrates that although the model generated ten topics, not all topics appeared equally. Some topics dominated discussion in the corpus, while others were more specific or minor. This imbalance actually reflects the natural structure of news documents which tend to focus on certain issues that are currently being hotly discussed.

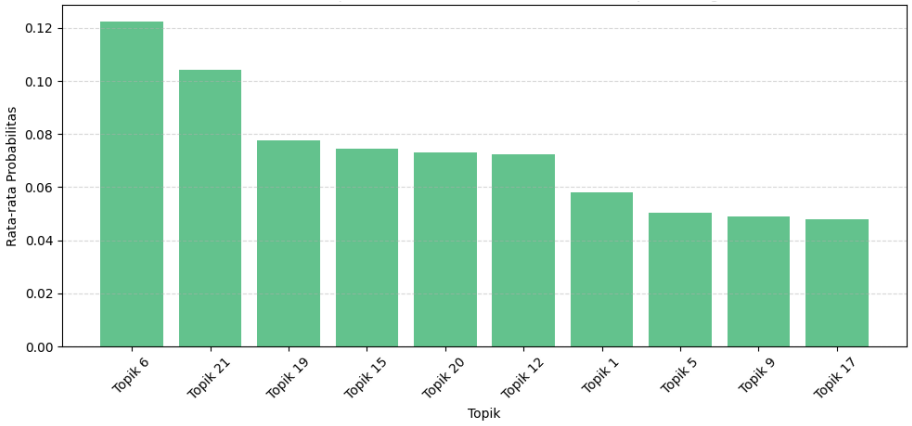

Figure 7. Topic Distribution Across Documents from Extended Text Preprocessing

Based on the topic and key words generated by the LDA model using data from extended text preprocessing, it appears that the resulting topics are more focused and less generic. For example, Topic 21 is dominated by words like "makan," " gizi," "sekolah," " siswa," " laksana," which explicitly refer to the implementation of free nutritious meals in schools. Meanwhile, Topic 6 contains the words "program," "makan," " gizi," and "terima," which indicate the operational or administrative aspects of program implementation. Visualizing the topic distribution across documents in the LDA model using data from extended text preprocessing (Figure 7) shows that the average probability of each topic is more selective, with certain topics (e.g., Topics 6 and 21) dominating with probability values above 0.1, while other topics have lower contributions. This indicates that documents in the extended model tend to have a clearer topic focus and are not evenly distributed across multiple topics, as is the case in the LDA model using data from basic text preprocessing. This contrasts with the LDA model for basic text preprocessing (Figure 6), where the topic distribution within the document appears more even. This suggests that without masking, the model struggles to distinguish meaning boundaries between phrases, resulting in overlapping topics and generic words like "program," "free," or "eat," spread across multiple topics.

3.4. Evaluation of Coherence Score and Cosine Similarity

In the basic text preprocessing approach, the highest coherence value was obtained at K=10 with a coherence score of 0.3932, while the next highest values were found at larger K values, such as K=48 (0.3834) and K=35 (0.3823). Conversely, in the extended text preprocessing using masking multiword expressions (MWE), the LDA model consistently produced higher coherence scores. The peak was reached at K=22 with a value of 0.5149, followed by K=20 (0.5061) and K=23 (0.5019). Table 8 shows the five highest coherence results and their optimal number of topics using two cleaned datasets.

Table 8. K Values with the Highest Coherence Scores

K Basic	Coherence Score	K Extended	Coherence Score
10	0.3932	22	0.5149
48	0.3834	20	0.5061
35	0.3823	23	0.5019
15	0.3804	21	0.5015
59	0.3797	6	0.4964

In addition to coherence evaluation, cosine similarity is also used to evaluate the semantic closeness between topics generated by the LDA model. Each topic is represented as a word distribution vector, and its distance to other topics is calculated using the cosine metric. High cosine similarity values indicate similar semantic representations between topics, which may indicate topic duplication.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Table 9 shows the five topic pairs with the highest cosine similarity values in the basic preprocessing and extended preprocessing models.

Table 9. Cosine Similarity Values

Model	Topic Pair	Cosine Similarity
Basic	Topik 6 dan Topik 7	0.8244
	Topik 2 dan Topik 9	0.7398
	Topik 2 dan Topik 6	0.7324
	Topik 5 dan Topik 9	0.7187
	Topik 2 dan Topik 7	0.7035
Extended	Topik 5 dan Topik 18	0.7406
	Topik 19 dan Topik 20	0.7395
	Topik 5 dan Topik 19	0.7323
	Topik 0 dan Topik 18	0.7100
	Topik 4 dan Topik 20	0.7056

Table 9 show that the LDA model with basic text preprocessing produces topic pairs with higher similarity, such as Topic 6 and Topic 7, which have a cosine similarity value of 0.8244. This indicates that the two topics are very similar semantically and may overlap. In contrast, the LDA model with extended text preprocessing achieved the highest cosine similarity value at 0.7406. This indicates that the resulting topics have more discrete semantic representations, enabling the model to distinguish topics more clearly and specifically. This result also supports the previous topic coherence and distribution evaluation, which showed an increase in topic coherence after multiword expression (MWE) masking.

3.5. The Effect of Multiword Expression Masking on Topic Modeling

Multiword Expression (MWE) masking is performed as part of the extended text preprocessing stage to maintain the unity of meaning of important phrases that frequently appear in documents. Phrases such as "makan gizi gratis" (free nutritious meal), "program utama" (superior program), or "kitchen makan gizi" (nutritious meal kitchen) have specific and coherent meanings, so if the words are separated, the contextual meaning can be distorted. Therefore, these phrases are combined into a single token with an underscore (e.g., "makan_bergi_gratis") so that they are treated as a single meaning unit by the LDA model [24].

The impact of MWE masking can be seen significantly in the increase in topic coherence values. In the LDA model with basic text preprocessing, the highest coherence value was only 0.3932, whereas after applying MWE masking, the coherence value increased to 0.5149. This improvement indicates that the resulting topics are more semantic, consistent, and relevant. Furthermore, changes are also visible in the topic distribution within the documents. The visualization results

show that the topic distribution in documents resulting from advanced preprocessing has become sharper, meaning that each document is more dominant in only one or two topics. This contrasts with the LDA model using basic text preprocessing, which tends to show a flatter and more diffuse topic distribution, indicating unclear semantic representation [25]. Multiword expression (MWE) masking not only improves topic coherence based on the coherence score but also strengthens thematic focus at the document level. This suggests that data cleaning, including multiword expression (MWE) masking, plays a significant role in improving the resulting topic representation.

3.6. Discussion

The analysis of the Free Nutritious Meal Program (MBG) news articles using topic modeling has revealed meaningful insights into how media constructs narratives around this national issue. By comparing basic and extended text preprocessing—particularly focusing on the application of Multiword Expression (MWE) masking—this discussion highlights key implications, methodological strengths, and potential limitations in the modeling process.

The application of MWE masking has significantly improved topic clarity and coherence in the LDA model. This is evident from the substantial increase in coherence scores—from 0.3932 in the basic model to 0.5149 in the extended model. This improvement underscores the importance of semantic preservation in the preprocessing phase, especially when analyzing domain-specific corpora like policy-related news. MWE masking helped retain the contextual meaning of phrases such as "makan gizi gratis" or "program utama," which would otherwise be distorted if treated as separate tokens. In public discourse, these phrases are often used as policy terms or slogans and carry implications that transcend their literal word-by-word meaning. By treating them as single tokens, the extended preprocessing ensures that the model understands these key expressions as unified concepts, thereby allowing more accurate topic grouping.

Another key observation is the improved separation between topics in the extended preprocessing model. Cosine similarity scores between topic pairs were lower in the extended model, suggesting reduced semantic overlap. This allows each topic to represent a more distinct area of discourse. For instance, some topics clearly focused on political stakeholders, such as "Prabowo," "presiden," and "menteri," while others concentrated on logistical or technical aspects like "menu," "kantin," or "dapur." The basic preprocessing model, by contrast, resulted in more overlapping themes, with generic keywords like "program," "makan," and "gratis" appearing across several topics. This diluted the semantic identity of each topic and reduced the model's ability to differentiate between nuanced sub-themes.

In practical terms, the results from the extended preprocessing model offer more actionable insights for stakeholders such as policymakers, educators, and nutrition experts. For example, the identification of geographically specific discussions (e.g., "Manggarai," "Sikka," "Karanganyar") allows for targeted analysis of program implementation across regions. Similarly, topics focusing on operational terms like "dapur," "sppg," and "menu protein" highlight the areas where public and media attention are concentrated, potentially signaling program bottlenecks or public interest points. This level of interpretability is critical when using topic modeling for policymaking support. Policymakers rely on clear, focused themes to inform decision ns; muddled or overlapping topics (as seen in the basic model) would hinder this process.

The findings reinforce the necessity of adopting a data-centric approach in natural language processing (NLP), especially in languages like Indonesian where affixation and reduplication are prevalent. Traditional preprocessing steps—though widely used—may not sufficiently preserve the richness of meaning embedded in local expressions. As shown in this study, the inclusion of steps such as MWE detection and entity preservation using tools like IndoBERT can yield significantly better results. Moreover, this research opens pathways for further exploration into the integration of Named Entity Recognition (NER) and more sophisticated contextual embeddings that could further enhance topic differentiation. While this study used Latent Dirichlet Allocation (LDA), integrating deep learning-based models like BERTopic with strong preprocessing foundations could produce even more nuanced insights.

While the extended preprocessing model has shown superior performance, it is not without limitations. The detection of multiword expressions relies heavily on the accuracy of the IndoBERT model, which may not capture all relevant expressions in a domain-specific corpus. Furthermore, the decision to mask certain expressions involves subjective judgment that could bias results if not carefully validated. Additionally, while coherence and cosine similarity are effective evaluation metrics, they may not fully capture the subjective quality of topic interpretability from a human reader's perspective. Future research could benefit from incorporating qualitative assessments, such as expert reviews or public opinion analysis, to validate the relevance and usefulness of identified topics.

The improved quality of topic modeling achieved through enhanced preprocessing also holds broader implications for how digital media narratives are analyzed. As media continues to shape public discourse on critical national programs like MBG, the ability to accurately extract dominant themes can help monitor misinformation, track sentiment shifts, and assess public engagement. Moreover, the use of topic modeling with robust preprocessing can be a valuable tool for government communication strategies. By understanding the evolving themes in public

narratives, communication teams can tailor their messaging, correct misunderstandings, and emphasize program strengths more effectively.

4. CONCLUSION

Based on the results of the analysis and experiments that have been conducted, it can be concluded that this study shows that masking multiword expressions (MWE) including entities, collocations, and compound words improves coherence and topic representation in Latent Dirichlet Allocation (LDA)-based modeling. The LDA model with extended text preprocessing (which involves masking) produces an optimal coherence score K of 0.5149, higher than the LDA model with basic text preprocessing of 0.3932 for its optimal coherence score K. The distribution of topics in documents in the LDA model with extended text preprocessing is sharper and more focused, indicating a more representative topic mapping of the document content. Cosine similarity analysis shows that topics in the LDA model with basic text preprocessing tend to be more similar to each other, while in the LDA model with extended text preprocessing, topics are more diverse and do not overlap. Masking MWE is proven to be able to maintain the relatedness of phrase meaning, reduce meaning distortion, and improve overall topic representation.

REFERENCES

[1] E. Setyawan, Rianto, Kusuma Wardana, Sugihartanto, Rizal Angko Pratama, and Malik Ibrahim, "Analisis Wacana Berita Hoaks tentang Program Makan Bergizi Gratis (MBG) Menggunakan Pendekatan Socio-Cognitive Teun A. van Dijk," *Jurnal Audiens*, vol. 6, no. 2, pp. 254–277, Jun. 2025, doi: 10.18196/jas.v6i2.607.

[2] D. Wulandari, N. Istiqomah, T. Utami, and Y. Sunesti, "Efektivitas Pengalokasian Dana Desa Terhadap Program Percepatan Penurunan Stunting," *Jurnal Pendidikan Sejarah dan Riset Sosial Humaniora (KAGANGA)*, vol. 7, no. 1, 2024.

[3] A. Santoso, B. D. Melianawati, and E. A. Ayuningtyas, "Governance Analysis Of The Implementation Of The Free Nutritious Meal Program," *Jurnal Manajemen Bisnis dan Organisasi (JMBO)*, vol. 4, no. 1, pp. 240–270, 2025, doi: 10.58290/jmbo.v4i1.423.

[4] A. Albaburrahim, A. P. A. Putikadyanto, A. N. Efendi, M. A. Alatas, S. Romadhon, and L. R. Wachidah, "Program Makan Bergizi Gratis: Analisis Kritis Transformasi Pendidikan Indonesia Menuju Generasi Emas 2045," *Entita: Jurnal Pendidikan Ilmu Pengetahuan Sosial dan Ilmu-Ilmu Sosial*, pp. 767–780, May 2025, doi: 10.19105/ejpis.v1i.19191.

[5] D. K. Geeganage, Y. Xu, and Y. Li, "A Semantics-enhanced Topic Modelling Technique: Semantic-LDA," *ACM Trans Knowl Discov Data*, vol. 18, no. 4, Feb. 2024, doi: 10.1145/3639409.

[6] T. Wada, Y. Matsumoto, T. Baldwin, and J. H. Lau, "Unsupervised Paraphrasing of Multiword Expressions," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.01443

[7] M. Jelita, "Text Mining dengan Topic Modelling LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI," *Media Pustakawan*, vol. 31, no. 3, pp. 253–265, Dec. 2023, doi: 10.37014/medpus.v31i3.5237.

[8] A. Breuer, "E-LDA: Toward Interpretable LDA Topic Models with Strong Guarantees in Logarithmic Parallel Time," Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.07747

[9] H. Sudarman, "Analisis dan Deteksi Kemiripan Teks Berbasis Python dengan Algoritma Levenshtein Distance," *Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)*, vol. 10, pp. 257–273, 2025.

[10] S. Sahoo, J.Maiti, and V.K.Tewari, "Multivariate Gaussian Topic Modelling: A novel approach to discover topics with greater semantic coherence," 2025.

[11] A. Amalia, O. Salim Sitompul, E. Budhiarti Nababan, and T. Mantoro, "A Comparison Study of Document Clustering Using Doc2vec Versus Tfidf Combined with Lsa for Small Corpora," *J Theor Appl Inf Technol*, vol. 15, p. 17, 2020.

[12] I. Zaitova, V. Hirak, B. M. Abdullah, D. Klakow, B. Möbius, and T. Avgustinova, "Attention on Multiword Expressions: A Multilingual Study of BERT-based Models with Regard to Idiomaticity and Microsyntax," May 2025. [Online]. Available: http://arxiv.org/abs/2505.06062

[13] H. Kresnawan, S. G. Felle, H. G. Mokay, and N. A. Rakhmawati, "Analyzing Main Topics Regarding the Electronic Information and Transaction Act in Instagram Using Latent Dirichlet Allocation," *Data Science: Journal of Computing and Applied Informatics*, vol. 5, no. 2, pp. 71–84, Jul. 2021, doi: 10.32734/jocai.v5.i2-6125.

[14] A. Drissi, S. Sassi, R. Chbeir, A. Tissaoui, and A. Jemai, "SemaTopic: A Framework for Semantic-Adaptive Probabilistic Topic Modeling," *Computers*, vol. 14, no. 9, Sep. 2025, doi: 10.3390/computers14090400.

[15] H. Mu, S. Zhang, and H. Xu, "A Knowledge-Driven Approach to Enhance Topic Modeling with Multi-Modal Representation Learning," *in ICMR 2024 - Proceedings of the 2024 International Conference on Multimedia Retrieval, Association for Computing Machinery, Inc*, May 2024, pp. 1347–1355. doi: 10.1145/3652583.3658069.

[16] B. Warsito, J. Endro Suseno, and A. Arifudin, "Embedding and Topic Modeling Techniques for Short Text Analysis on Social Media: A Systematic Literature Review," *Data and Metadata*, vol. 4, p. 1168, Sep. 2025, doi: 10.56294/dm20251168.

[17] J. Schneider, "Efficient and Flexible Topic Modeling Using Pretrained Embeddings and Bag of Sentences," *in International Conference on Agents and Artificial Intelligence, Science and Technology Publications*, *Lda*, 2024, pp. 407–418. doi: 10.5220/0012404000003636.

[18] H. Sakai and S. S. Lam, "HAMLET: Healthcare-focused Adaptive Multilingual Learning Embedding-based Topic Modeling," 2025.

[19] T. P. Nguyen et al., "XTRA: Cross-Lingual Topic Modeling with Topic and Representation Alignments," Oct. 2025. [Online]. Available: http://arxiv.org/abs/2510.02788

[20] G. Kumar Das and P. Bhattacharjee, "eLDA: Augmenting Topic Modeling with Word Embeddings for Enhanced Coherence and Interpretability," *Journal of Information Systems Engineering and Management*, vol. 2025, no. 21s, pp. 2468–4376, 2024.

[21] Y. Kustiyahningsih and Y. Permana, "Penggunaan Latent Dirichlet Allocation (LDA) dan Support-Vector Machine (SVM) Untuk Menganalisis Sentimen Berdasarkan Aspek Dalam Ulasan Aplikasi EdLink," *Teknika*, vol. 13, no. 1, pp. 127–136, Mar. 2024, doi: 10.34148/teknika.v13i1.746.

[22] A. Yaman, B. Sartono, A. M. Soleh, and I. Pertanian Bogor, "Pemodelan topik pada dokumen paten terkait pupuk di Indonesia berbasis Latent Dirichlet Allocation 1 2 3," *Berkala Ilmu Perpustakaan dan Informasi*, vol. 17, no. 2, pp. 168–180, 2021, doi: 10.22146/bip.v17i1.2147.

[23] Kristine Angelina Simanjuntak, Muhamad Koyimatu, Yolla Putri Ervanisari, and Tasmi, "Identifikasi Opini Publik Terhadap Kendaraan Listrik dari Data Komentar YouTube: Pemodelan Topik Menggunakan BERTopic," *TEMATIK*, vol. 11, no. 2, pp. 195–203, Dec. 2024, doi: 10.38204/tematik.v11i2.2096.

[24] L. Nur Halimah, S. Riyadi, A. Fatahillah Jurjani, A. Prayogi, and S. Dwi Laksana, "Implementasi Penggunaan Machine Learning Dalam Pembelajaran: Suatu Telaah Deskriptif," *Journal Penelitian Pendidikan*, vol. 1, no. 1, 2025.

[25] D. Mubarok et al., "Big Data Analytics Dan Machine Learning Untuk Memprediksi Perilaku Konsumen Di E-Commerce," *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 8, no. 1, 2025.