GENERATION JOURNAL

Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

Implementation of Topic modeling for Multilingual Document Summarization based on Bag of Itemset

Bambang Subeno [1]
[1] School of Computing, Telkom University, Jl. Telekomunikasi No.1 Bandung, Indonesia
*corresponding author : bambang.subeno.if@gmail.com

Article Info

Received	2026-02-02
Revised	2026-03-13
Accepted	2026-03-14

Keywords : Maximum 5 keywords, separate with a comma(;)

Abstract

With the increasing number of electronic text documents, the process of searching and processing information has become increasingly complex, especially when these documents come from multiple sources and languages. Consequently, document summarization methods are needed to help users retrieve important information more quickly. However, existing multilingual summarization methods, such as ELSA, are limited by dataset size and the need to pre-determine themes. By integrating the Bag of Itemset representation and the Latent Dirichlet Allocatio Algorithm Modification (LDA-AM) approach, this study aims to improve the quality of multilingual document summarization. The proposed method first uses topic modeling to divide different multilingual documents into several topics. Then, for each topic, a sentence selection process is performed to generate topic-based summaries, which are then combined into a general summary. Using the ROUGE evaluation metric, experiments were conducted to compare the proposed method with baseline. Experimental results show that the proposed method performs better than ROUGE-1 with a value of 0.2623, ROUGE-2 with a value of 0.1802, and ROUGE-L with a value of 0.1231. The results indicate that in the process of summarizing multilingual documents, summary quality can be improved by combining the Bag of Itemset representation and LDA-AM.

INTRODUCTION

Along with the development of the internet and big data, electronic textual documents are increasingly used, easy to access, and increasingly common in several application contexts. Examples include digital libraries, online news, social media, and e-learning platforms. The diffusion of electronic textual documents has opened up a number of research branches, including text summarization, document indexing, retrieval, and visualization. With the rapid development of textual data spread across the internet, people are overwhelmed by the amount of information and documents on the internet. This has triggered the desire of many researchers to develop a technological approach that can automatically summarize text that produces a summary containing important sentences and includes all relevant important information from the original document [1].

Summarization is a collection of short texts that should represent most of the information in the source text and cover most of its topics [2,3]. Summarization is needed to support data analysis, and summarization allows effective use for exploration of large data sets [4]. Research on text summarization has been studied since the mid-20th century; the first summarization technique used was word frequency diagrams [5]. Until now, there have been many summarization studies, both with the

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

approach of the number of documents, namely single and multi-documents, and based on the summary of the results, namely extractive, abstractive, and Hybrid [6].

Multiple-document summarization is a summarization that automatically produces short summaries from large collections of textual documents [7,8], making it easier for readers to read the entire contents of the document, increasing the accessibility of content in the context of applications characterized by bandwidth limitations or visualization problems. Research that has been done effectively produces summaries consisting of subsets of document sentences. Research that has been done relies on linguistic models (ontologies, lexical databases) to obtain semantics from text or on machine learning and data mining algorithms to obtain significant correlations between text data. Research on linguistic models for different languages is still limited; in recent years, many research efforts have been made specifically to propose multilingual summarization algorithms based on machine learning [6,9].

Multilingual summarization is a summarization consisting of a collection of documents in different languages. Unlike cross-language, multilingual documents in the same collection are all written in the same language [9]. Many methods are used in text summarization, including Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF), RIPTIDES, neural networks, Feed Forward Neural Network (FFNN), IncreSTS, RBP-SUM [5], TextRank [10], and ELSA Summarizer [9]. ELSA Summarization is a multilingual summarization method that combines the LSA and Itemset Frequency methods [9]. Multilingual summarization using the ELSA Summarization algorithm is still limited to small data, and input data is determined homogeneously at the beginning [9]. In this study, document summarization will be carried out using a topic summarization approach on documents using the Bag of Itemset and LDA-AM (Latent Dirichlet Allocation-Algorithm Modification) approaches, so that data input can use heterogeneous data. Documents will be grouped into the same topic, then summarized for each topic so that it will form a multi-summarization based on the topics formed. The primary contribution of this study is the integration of the Bag of Itemset representation with the LDA-AM topic model to enable the multilingual, topic-based collection of heterogeneous texts utilized in summarization.

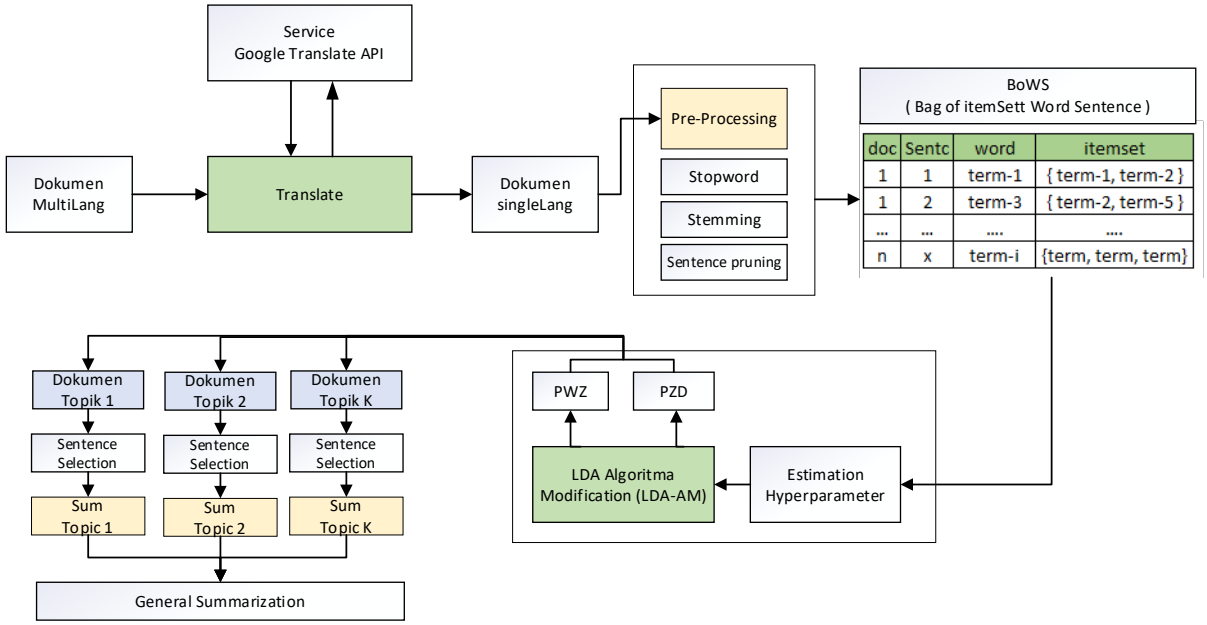

Figure 1. Topic Modeling Framework Model for Multilingual Document Summarization

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

RESEARCH METHOD

In this study, multilingual document summarization will be carried out according to the summarization framework in Figure 1. This framework illustrates the entire process of multilingual document summarization, which includes translating multiple languages, preprocessing, topic modeling using LDA-AM, sentence selection, and generate summarization.

Based on Figure 1, the summarization process that will be carried out consists of several stages, namely the document translation process, pre-processing, topic modeling, sentence selection, and generating summarization. The dataset used in this study is the MultiLing 2015 Multilingual Single-Document Summarization (MSS) dataset, which was introduced at the MultiLing workshop in SIGDIAL 2015 [10]. This dataset consists of Wikipedia articles in various languages, namely English, Italian, German, Spanish, and French. Statistics related to the Multiling 2015 dataset can be seen in Table 1.

Table 1. Statistics of the Multiling 2015 Dataset

Language	Average Doc	Words	Vocabulary
English	30	1,890,356,976	973,839
Italian	30	371,218,773	378,286
German	30	657,234,125	1,042,683
Spanish	30	464,465,399	419,683
French	30	551,057,299	458,748

Translate Documents

At this section, the multilingual input document will be translated into English first using the Google Translate API service. So that the document will be formed in one language. After the single-language document is formed, it will be processed to the next stage, namely the pre-processing stage.

Preprocessing

At this section, documents that are already in single-language form will be pre-processed first. The pre-processing stages carried out are (1) Stopword elimination, at this stage words that have little lexical content are removed. (2) Stemming, at this stage words that appear in sentences are changed into their basic form. (3) Sentence pruning, this stage considers the first sentence in each document, removing short sentences. After pre-processing is complete, a bag of itemset will be formed. This bag of itemset contains words that are grouped into itemsets. The representation of the bag of itemset can be seen in Figure 2.

All tables and figures must have names and numbers. All tables and figures must be referenced by number in the explanation or description. Examples of the format for writing figures and tables can be seen in Figure 1 and Table 1. The text in the table uses a font size of 10.

doc	Sentc	word	itemset
1	1	term-1	{ term-1, term-2 }
1	2	term-3	{ term-2, term-5 }
...	...	....	....
n	x	term-i	{term, term, term}

Figure 2. Bag of Itemset Representation

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

Topic Modeling

At this stage, the process of grouping documents based on the same topic will be carried out using LDA-AM. LDA-AM is a modified LDA model with an approach to eliminating document looping [11]. LDA is a probabilistic generative model that represents documents as a mixture of multiple latent topics, where each topic is represented by a probability distribution over words or itemsets. The model estimates two main probability distributions: the document-topic distribution and the topic-itemset distribution, which represent the probability of a topic appearing in a document and the probability of an itemset appearing in a topic, respectively.

The document-topic distribution indicates the probability of a topic appearing in a given document, while the topic-itemset distribution represents the probability of an itemset appearing in a given topic. The estimation of both probability distributions is performed using the Gibbs Sampling method. In this process, each itemset in the document collection is iteratively assigned a topic label based on a probability distribution that is repeatedly updated until convergence is achieved. The primary goal of LDA-AM is to minimize the use of duplicate documents during the Gibbs Sampling process and to identify relationships between adjacent words, which cannot be captured by the conventional bag-of-words approach. This objective is realized through the utilization of a Bag of Itemset Word Sentence (BoWS) representation, which is prepared prior to the commencement of the data processing stage.The LDA-AM pseudocode can be seen in Figure 4, and the classic LDA pseudocode can be seen in Figure 3.

for (d=1 to D) **do**
  for (i=1 to N_d) **do**
    $v \leftarrow w_{di}$, $I_{di} \leftarrow N_{di}$
    for (j = 1 to I_{di}) **do**
      k = z_{dij}
      $N_{wk} \leftarrow N_{wk} - 1$, N$N_{dk} \leftarrow N_{dk} - 1$
      for (k = 1 to K) **do**
        $pk = (N_{wk} + \beta)$ x $(N_{dk} + \alpha)$ / $(\sum_v N_{vk} + V\beta)$
      $x \sim uniform\ (0, pk)$
      k $\leftarrow binarySearch(k: p_{k-1} < x < p_k)$
      $N_{wk} \leftarrow N_{wk} + 1$, N$N_{dk} \leftarrow N_{dk} + 1$
      z_{dij} = k

Figure 3. Classic LDA pseudo-code [12]

As illustrated in Figure 3, the pseudocode of classical LDA employs the Gibbs Sampling method to estimate topic distributions. The process commences with the iteration of each document d in the corpus, followed by the subsequent iteration of each word w contained within that document. For each word, the temporary topic assignment $zdij$ is first removed by subtracting the counter values from the word-topic (Nwk) and document-topic (Ndk) distributions. In the following step, the probability of each topic pk is calculated on the basis of the combination of the word-topic and document-topic distributions. These distributions have been refined using the hyperparameters α and β. Subsequent to the acquisition of the topic probabilities, new topics are selected at random based on these probability distributions, and the counter values are then updated once more. This process is repeated iteratively for all documents until the topic distribution reaches a stable state.

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

Pre-processing ← indexing word dokumen (N_{di})

for (i=1 to N_d) **do**
  $v \leftarrow w_{di}$, $I_{di} \leftarrow N_{di}$
  for (j = 1 to I_{di}) **do**
    k = z_{dij}
    $N_{wk} \leftarrow N_{wk} - 1$, N$N_{dk} \leftarrow N_{dk} - 1$
    for (k = 1 to K) **do**
      $pk = (N_{wk} + \beta)$ x $(N_{dk} + \alpha)$ / $(\sum_v N_{vk} + V\beta)$
      $x \sim uniform\ (0, pk)$
      k ← $binarySearch(k: p_{k-1} < x < p_k)$
      $N_{wk} \leftarrow N_{wk} + 1$, N$N_{dk} \leftarrow N_{dk} + 1$
      z_{dij} = k

Figure 4. LDA-AM pseudo-code

As illustrated in Figure 4, the pseudocode of the LDA-AM approach, a modification of the classic LDA, is presented. The primary distinction of this approach is the elimination of document-based iterations, which are substituted for a vocabulary-based or itemset-based approach. This approach is obtained in the preprocessing stage using the Bag of Itemset Word Sentence (BoWS) representation. In this approach, each item is processed based on the vocabulary index, thereby ensuring that documents containing the same words or patterns do not need to be processed repeatedly. The topic assignment process continues to be executed through the utilization of the Gibbs Sampling mechanism. The LDA-AM process runs once inference will produce the probability value of the itemset topic for the tth itemset and k topics (φk,t = PWZ) according to the equation (1) and the proportion value of the topic in document d (θd,k = PZD) according to the equation (2).

$$\varphi_{k,t} = PWZ = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_k^{(t)} + \beta_t)} \quad (1)$$

$$\theta_{d,k} = PZD = \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_d^{(k)} + \alpha_k)} \quad (2)$$

Sentence Selection

At this section, sentence selection will be carried out to select sentences that will be used as summary candidates based on the value of the item set-topic probability (PWZ). PWZ represents the probability of an itemset Wgiven a topic Z, which corresponds to the topic–itemset distribution [13]. This distribution indicates how strongly an itemset is associated with a particular topic and is used to determine the importance of sentences containing the itemset. Sentences that will be used as topic summary candidates are three sentences that have the highest item set-topic probability results.

Generate Summarization

At this section, the generate summarization process is carried out. The generate summarization process is carried out based on each topic; each topic will produce one summary. The summary on the topic is extractive, which takes from sentences contained in the topic document collection.

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

RESULTS AND DISCUSSION

At this stage, two experimental scenario processes are carried out, namely the first is summarizing based on the proposed model framework, then comparing the summarization results with the baseline method. In this scenario, the quality of the summary was evaluated using the ROUGE metric by comparing the summaries generated by the system with the reference summary. Since the proposed model framework produces a summary for each identified topic, the ROUGE value was calculated for each topic summary, and then the final ROUGE value was obtained by calculating the average ROUGE value across all topics. In previous studies using the ELSA method, evaluations were only reported using the ROUGE-2 metric, for ROUGE-1 and ROUGE-L scores were not reported in the original publication. Therefore, in this study, the ELSA method was rerun on the same dataset to obtain ROUGE-1 and ROUGE-L scores, allowing for a more comprehensive evaluation of summarization performance. The second is comparing the computation time of running summarization between the proposed model and baseline. The results of the first scenario can be seen in Table 2, and the results of the second scenario can be seen in Figure 5.

Table 2. Comparison of Result Summarization Method

Summarization Method	ROUGE-1	ROUGE-2	ROUGE-L
ELSA	0,2514	0,1742	0,1025
TextRank	0,1918	0,1006	0.0935
Centroid	0,2312	0,0991	0.0906
Our Model	**0,2623**	**0,1802**	**0,1231**

Based on table 2, the results of the summarization evaluation using ROUGE-2 show that the proposed method gets better results, namely 0.1802. These results show better results than the baseline methods ELSA, TextRank, and Centroid. This shows that the use of a bag of itemset and LDA-AM can improve summarization results.

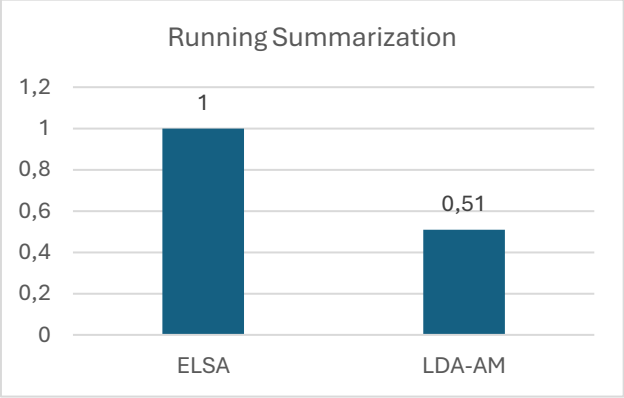

Figure 5. Summarization Computation Time

Based on Figure 5, the computation time of using LDA-AM is faster to perform the summary process compared to the ELSA summarization method. The ELSA method requires approximately 49% more computation time than the LDA-AM method. This increased efficiency is due to the fact that the LDA-AM method processes vocabulary-based data and eliminate document looping, thereby reducing the redundancy of processing the same word across multiple documents. Thus, the summarization process can be performed more efficiently than methods that still iterate directly on each document. This

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

shows that the proposed method, in addition to better summary results in terms of computation time, also shows faster results.

CONCLUSION

Based on the experimental results, the proposed summarization method, which integrates the Bag of Itemset Word Sentence (BoWS) representation and the LDA-AM approach, can improve document summarization performance. Evaluation results show that the proposed model achieved ROUGE-1 scores of 0.2623, ROUGE-2 scores of 0.1802, and ROUGE-L scores of 0.1231, which are higher than the baseline methods ELSA, TextRank, and Centroid. This indicates that the use of the itemset representation is able to better capture the relationships between words in a sentence, resulting in a more representative summary. In addition to improving summary quality, the proposed method also demonstrates better computational efficiency than the baseline method. This is achieved through the application of LDA-AM, which reduces processing redundancy by utilizing a vocabulary-based representation, thus minimizing document iteration. The main contribution of this research is the development of a document summarization method that combines the Bag of Itemset Word Sentence (BoWS) representation and LDA-AM topic modeling, which has been shown to improve both summary quality and computational efficiency. Further research could involve using deep learning-based approaches, such as BERT, to enhance semantic understanding in the summarization process. The BART model could also be employed to generate a more coherent, general summary based on the summaries of each topic. Furthermore, other topic modeling methods, such as BERTopic, could be explored to enhance topic representation in multilingual document summarization.

REFERENCES

.

[1] Y. Zang, H. Jein, D. Meng "A comprehensive survey on automatic text summarization with exploration of LLM-based methods," *Neurocomputing,*vol. 663, 2026.

[2] Jiang M, Zou Y, Zhang, "GATSum: Graph-Based Topic-Aware Abstract Text Summarization," *Information Technology and Control,* pp. 345-355, 2022.

[3] Meiling Xu, Hayati, "Text Summarization: A Bibliometric Study and Systematic Literature Review," *Ingénierie des Systèmes d'Information,* pp. 2207-2089, 2024.

[4] M. G. S. R. Matteo Francia, "Summarization and visualization of multi-level and multi-dimensional itemsets," *Information Sciences,* pp. 63-85, 2020.

[5] S. R. G. F. S. E. N. Adhika Pramita Widyasari, "Review of automatic text summarization techniques & methods," *Journal of King Saud University – Computer and Information Sciences,* pp. 1029-1046, 2020.

[6] Mr. Pranav, Mr. Shivprasad, Mr. Aniket, "Multilingual Text Summarization Using NLP," International Journal of Advanced Research in Science, Communication and Technology(IJARSCT), vol. 5, 2025.

GENERATION JOURNAL
Vol. 10 No. 2 Tahun 2026
e-ISSN: 2549-2233 / p-ISSN: 2580-4952
DOI : https://doi.org/10.29407/gj.v10i2.28393

[7] Y. W. W. S. X. Z. A. R. a. Y. Y. X. Yan, "Unsupervised Graph-Based Tibetan Multi-Document Summarization," *Computers, Materials and Continua,* pp. 1769-1781, 2022.

[8] S. B.-G. a. Z. Z. D. R. Radev, "Experiments in Single and Multi-Document Summarization Using MEAD," *The First Document Understanding Conference,* 2021.

[9] P. G. a. E. B. P. d. T. Luca Cagliero, "ELSA: A Multilingual Document Summarization Algorithm Based on Frequent Itemsets and Latent Semantic Analysis," *ACM Transactions on Information Systems,* vol. 37, 2019.

[10] R. Gaetano, B. Pierpaolo, S. Giovanni, "Centroid-based Text Summarization through Compositionality of Word Embeddings," *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pp. 12-21, 201, 2017.

[11] B. Subeno, R. Kusumaningrum dan Farikhin, "Optimisation towards Latent Dirichlet Allocation: Its Topic Number and Collapsed Gibbs Sampling Inference Process," *International Journal of Electrical and Computer Engineering (IJECE),* pp. 3204-3213, 2018.

[12] T. S. X. Han, "Efficient Collapsed Gibbs Sampling For Latent Dirichlet Allocation," *Asian Conference on Machine Learning (ACML2010),* 2010.

[13] R. Dani, W. Deden, S. Dady, "Observing the Performance of the TextRank Algorithm on Automatic Text Summarization for Bahasa Indonesia," *International Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, pp. 1147–1153, 2023.