

Terbit online pada laman : <http://teknosi.fti.unand.ac.id/>

## Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |



Research Article

# Hybrid LexRank-LDA-MMR for Indonesian Text Summarization

Nasrul Amin Muis<sup>a,\*</sup>, Yoga Pristyanto<sup>b</sup>, Ika Nur Fajri<sup>c</sup>

<sup>a,b,c</sup> Information Systems Study Program, Faculty of Computer Science, Amikom University Yogyakarta, Indonesia.

### ARTICLE INFORMATION

*Sejarah Artikel:*

Diterima Redaksi: 16 Mei 2025

Revisi Akhir: 27 April 2026

Diterbitkan Online: 05 Mei 2026

### KEYWORDS

Extractive Summarization,  
LexRank,  
LDA,  
MMR,  
Hybrid Approach,  
ROUGE

### CORRESPONDENCE

E-mail: [nasrulinuis68@gmail.com](mailto:nasrulinuis68@gmail.com)\*

### ABSTRACT

The rapid growth of digital text information makes it crystal clear that there is a need for automated tools that summarize text for rapid retrieval. Extractive methods employed include LexRank, Latent Dirichlet Allocation (LDA), and Maximal Marginal Relevance (MMR), and the study aimed to enhance the quality of Indonesian text summaries beyond regular LexRank. In this study, the role of LexRank was to assist in selecting meaningful sentences that were centric to the center of the graphs, while the role of LDA was to ensure that the sentences were typically relevant. The strength of MMR lies in maintaining the document's relevance and diversity, thereby reducing redundancy in the summaries. Summaries from two publicly available datasets, IndoSum and Liputan6, containing texts in Bahasa Indonesia, were analyzed at 30% and 50% compression levels and graded using ROUGE (ROUGE-1, ROUGE-2, ROUGE-L F1) scores. Analysis of 5000 articles per dataset showed that implementing LexRank and LDA together with MMR resulted in a higher average ROUGE score than standard LexRank, irrespective of the set compression levels and across both datasets, demonstrating the approach's effectiveness in enhancing summary quality. The improvements recorded are most significant in ROUGE-1 and ROUGE-2, indicating that these combination approaches can produce more informative and relevant summaries while preserving sentence-level diversity, thereby deepening understanding of the information presented in the summary.

## 1. INTRODUCTION

One significant effect of the digital era is the tremendous growth of textual data. To illustrate, internet users generate nearly 2.5 quintillion bytes of data daily [1]. This poses several problems, especially when finding important information in online news articles, papers, social media, and other documents. How do we manage everything and make sense of it? The problem is where Automatic Text Summarization (ATS) steps in as an essential answer. The goal is to produce concise representations of one or more documents while preserving key information from the original documents. Generally, text summarization techniques have two main types: extractive and abstractive [2]. In the extractive method, significant sentences from the source document are selected and put together to form a summary. The abstractive method, on the other hand, tries to understand the

essence of the text. The summary generator creates the summary differently from how the source document presents it. Although more coherent and succinct, the stronger summaries generated using abstractive methods tend to be much more challenging to build, require large amounts of training data and computation resources, and are much less efficient [3]. On the other hand, solely extractive techniques tend to preserve original sentences, thereby increasing the likelihood of yielding accurate summaries while requiring less computational work. A well-known extractive ranking method is LexRank. It applies the PageRank algorithm, originally developed to measure the importance of web pages. In LexRank, a graph represents the text, with sentences as nodes and connections between nodes as edges. The algorithm determines each sentence's relationship based on the cosine similarity of the TF-IDF representation. Despite this usefulness, traditional LexRank has some shortcomings. Previous research has shown that standard lexrank may include sentences

that do not revolve around the document's main subject or that are similar or redundant, because it relies solely on centrality [4]. The statements above highlight key points: extraction-based summarization methods have limitations, while generation-based techniques rely on underlying intricacies. The other side of the coin states precisely that. Therefore, we need to explore strategies of a different nature. To overcome these limitations, this research introduces a hybrid approach that adds two additional methods to LexRank: LDA and MMR. LDA is a model within the family of probabilistic topic models and helps determine the presence and distribution of topics in a document corpus [5]. LDA assesses whether a sentence relates to the document's key themes. Conversely, MMR is a technique specifically designed to fetch relevant items (sentences, in this case) based on a query score, while ensuring the items are as distinct as possible from the previously selected set [6]. The expectation for combining these three methods is to produce a summary that includes central sentences selected by LexRank, thematically relevant sentences provided by LDA, and non-redundant sentences selected by MMR. This study focused on determining whether combining LexRank with LDA and MMR, while preserving the three-part structure, yields higher-quality summaries than a single LexRank system, particularly for Indonesian texts. This analysis will focus on the IndoSum and Liputan6 datasets, using ROUGE evaluation metrics to assess the summaries produced. The promise this hybrid method holds for considerably enhancing the accuracy of summarization for Indonesian-language texts is a very encouraging prospect for the future of natural language processing (NLP) [7].

## 2. METHOD

The research methodology is systematic and begins with data preparation. The underpinning dataset should always maintain relevance and quality enough to back the credibility of the analysis [8]. Later in the document, a phase is described as data sanitization, which involves removing unhelpful data to maintain the informativeness of the dataset. The standard LexRank method executes the ranking algorithm, making it the most efficient method for extracting essential information within the text. Other methods, such as LexRank+LDA+MMR, are also applied, but they aim to improve ranking results. In addition to conducting qualitative analysis, we used ROUGE scores to assess the quality of the summary. The assessed metric framework calculates a ranking score for a reference, which constitutes one of the evaluated indicators of an algorithm's performance. It ensures that all computations geared toward the algorithm-constrained scoring algorithm goals are reference-score resolutions. The researchers considered bounded constraints in their attempts to derive these metrics and to authenticate the results [9].

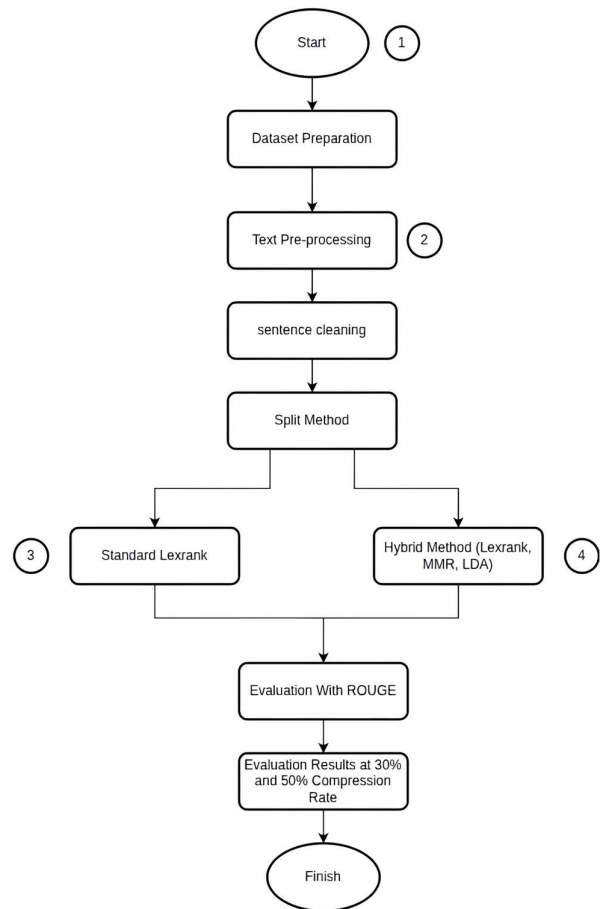


Figure 1. Research method

### 2.1. Dataset

The work used two publicly available Indonesian-language text summarization datasets. The datasets are essential for natural language processing in two ways: first, there is IndoSUM, a dataset of documents and summaries scraped from Indonesian online news portals. Enabling scholars to assess model outcomes is "This dataset consolidates data about a population.". Second, the Liputan6 news portal collected the Liputan6 dataset. This dataset also offers matched pairs of news articles with corresponding summaries, which enables further analysis of current news's evolving constructs and contents." For the evaluation phase, we randomly selected 5000 sample articles from the training split of each dataset. The aim is to automate the generation of performance metrics. Implementing training splits for this evaluation stems from the need for summary references that are vital for detailed analysis. No matter what, it is helpful to mention that in the review of practical applications, self-evaluations should, as previously discussed in the literature, be performed on split held-out validations to enhance model generalizability and reduce evaluative bias [10].

### 2.2. Text Preprocessing

The most fundamental initial steps in this process of operations are understanding the original document context with the intent of information retrieval, including detailing document cleansing.

For instance, this step applies to the NLTK library, which can break the text into sentence units. As it is known, this library is a central tool for working with English text in Natural Language Processing (NLP). It is important to note that simple Bahasa Indonesian sentences work well with NLTK. There is evidence that Indonesian-tailored tokenizers perform better in analyzing the text. Therefore, one must pay close attention when choosing the specific tokenization technique for the preprocessing stage. Moreover, we usually omit extremely short (10 characters or fewer) and empty sentences after the post-cleaning process, as they contribute little to a comprehensive analysis. Following these practices guarantees that subsequent data processing will be pertinent and high-quality [11].

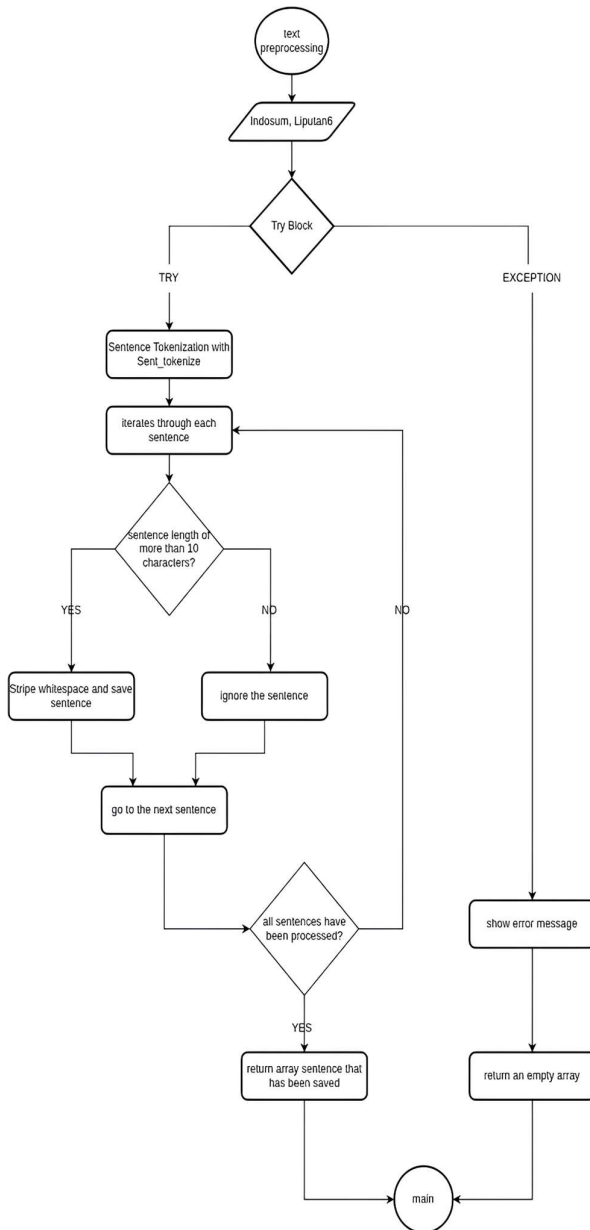


Figure 2. Text preprocessing flowchart

### 2.3. Summarization Method

Two techniques for extractive summarization are executed and assessed. The algorithm performs the LexRank process as follows: Sentence Representation: Each sentence in a document is represented as a vector using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF allocates weight to words based on their occurrence in the sentence and their infrequency across other sentences in the document. Similarity Matrix: Cosine similarity is computed for each pair of TF-IDF sentence vectors to create a similarity matrix. The  $S_{i,j}$  value indicates the level of similarity between sentence  $i$  and sentence  $j$ , where  $S_i$  and  $S_j$  represent the  $i$  and  $j$  sentences,

$$Sim(S_i, S_j) = \frac{TFIDF(S_i) \cdot TFIDF(S_j)}{\|TFIDF(S_i)\| \cdot \|TFIDF(S_j)\|} \tag{1}$$

and TF IDF (S) denotes the TF-IDF vector representation of sentence S. Calculation of the Centrality Score: The LexRank score for every sentence is determined utilizing the Power Method algorithm, which continuously updates the score based on the scores of its neighboring connected sentences in the similarity graph. The algorithm is comparable to PageRank.

$$p(u) = \frac{1-d}{N} + d \sum_{v \in In(u)} \frac{p(v)}{deg(v)} \tag{2}$$

Where  $p(u)$  signifies the LexRank score of the sentence  $u$ ,  $d$  is the damping factor (set to 0.85),  $N$  is the total count of sentences,  $v$  is the sentence linked to  $u$ , and  $deg(v)$  represents the number of connections from  $v$  (vertex degrees) [12]. The code implements the Power Method on its column-normalized similarity matrix.

$$s = \text{PowerMethod}(M, \alpha, \text{max\_iter}, \epsilon) \tag{3}$$

Sentence Selection: We organize sentences in descending order based on their LexRank score. We select the top  $k$  sentences to create a summary, where  $k$  is determined by the desired compression rate (compression\_rate) [13].

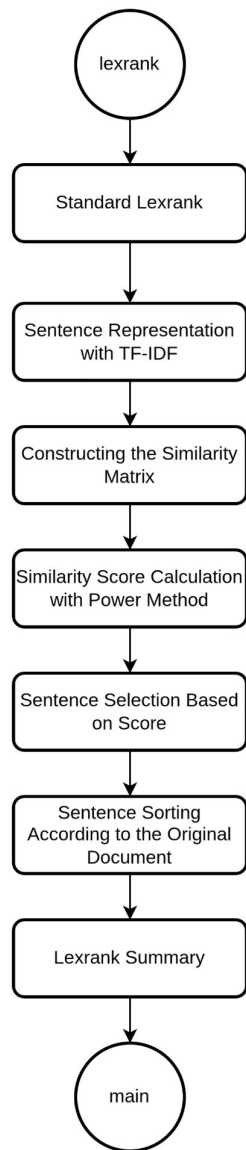


Figure 3. Standard lexrank flowchart

The system reorders the selected sentences according to their original appearance in the document using a combination of LexRank, LDA, and MMR. This approach merges the LexRank score with the LDA model's topic relevance and the MMR algorithm's diversity mechanism. The figure illustrates the process. The steps are: Calculate LexRank Scores: LexRank centrality scores (*lexrank\_scores*) are computed for all sentences using the standard method (steps 1-3 ). Calculate LDA Topic Relevance: The CountVectorizer (Bag-of-Words) model trains the LDA model utilizing a collection of sentences within the document. The number of topics (*LDA\_N\_TOPICS*) is 5 [14]. The trained LDA model derives the topic distribution for each sentence (*sentence\_topic\_dist*). The topic relevance score (*lda\_relevance\_scores*) indicates the probability that each sentence belongs to one of N topics. Combination and Normalization of Scores: LexRank and LDA relevance scores are normalized to a range of 0.1 using Min-Max Scaling. We

calculate the combined score (*combined\_scores*) by taking the weighted average of the two normalized scores. We assign weights (*alpha*) of 0.6 for LexRank scores and 0.4 (1-alpha) for LDA scores.

$$\begin{aligned}
 \hat{s}_{lexrank} &= \text{normalize}(s_{lexrank}) \\
 \hat{s}_{lda} &= \text{normalize}(s_{lda}) \\
 \alpha &= 0.6 \\
 s_{combined} &= \alpha \cdot \hat{s}_{lexrank} + (1 - \alpha) \cdot \hat{s}_{lda}
 \end{aligned}
 \tag{4}$$

Sentence selection using MMR: The Maximum Marginal Relevance (MMR) algorithm is used to select summary sentences. The MMR iteratively selects the sentence that maximizes the following score:

$$\text{MMR}(S_i) = \lambda \cdot \text{Relevance}(S_i) - (1 - \lambda) \cdot \max_{S_j \in S_{\text{selected}}} \text{Similarity}(S_i, S_j)
 \tag{5}$$

where *SI* is the candidate sentence [15]. *Relevance (Si)* is the relevance score of the *Si* sentence. In this case, *combined\_scores<sub>i</sub>.s\_selected* is the sentences already included in the summary. *Similarity (Si, Sj)* denotes the cosine similarity between the *Si* sentence and the already selected *Sj* sentence (derived from the initial similarity matrix calculated for LexRank). *λ (lambda\_param)* is the parameter that manages the balance between relevance and diversity (set to 0.3 in this experiment, favoring diversity). The selection process continues until we achieve the desired number of sentences (*k*). The algorithm rearranges the selected sentences in their original sequence.

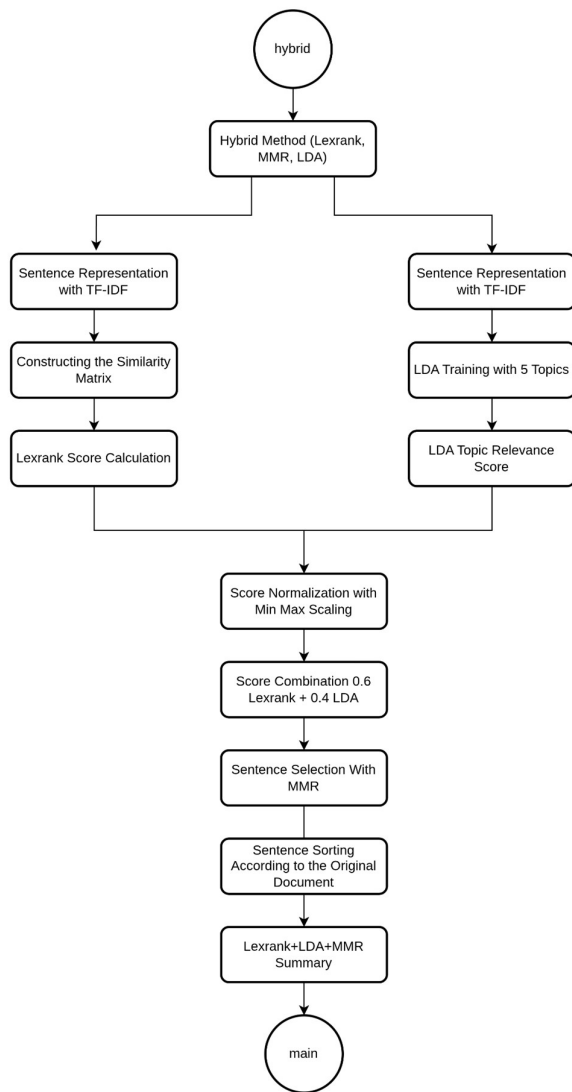


Figure 4. LexRank, lda, and mmr combined method flowchart

### 2.4. Evaluation

We used an automated evaluation system, ROUGE, to assess the quality of summaries generated by both methods. In particular, ROUGE-N (n-gram overlap) and ROUGE-L (longest subsequence) are the most frequently used ones, as ROUGE serves as the standard for automated text summary evaluation. The F1-score metric measures how closely the summary matches the human summary. For example, with ROUGE-1, 1-grams overlap is calculated between the summary and the reference, determining how much information in the summary is included in the reference. Then there is ROUGE-2, which assesses the overlap of 2-word phrases (bigrams), suggesting more complex contextual relationships within the text. ROUGE-L measures assess the resemblance based on the longest subsequence, which permits somewhat deeper linguistic analysis. We evaluated the initial sentence's 30% and 50% compression levels, and the results verified the method's efficacy in preserving context during summarization. Across the 5,000 articles, the average ROUGE score was computed for each method, dataset, and compression

ratio, providing a fair assessment of how well each technique achieved a balance between precision and informativeness in summary creation.

### 2.5. Configuration Parameters

The table presents factors which, as these experiments show, are essential for correctly understanding the collected data. The table illustrates the Experimental Configuration Parameters, which form the basis for systematic inquiries and their reproducibility in science. Following associated measures improves the relevance of the metrics captured and the data collected. To track the variable in interaction within the experimental setup, we need to understand the parameters in their operational, functional, and working states. In this regard, appropriate parameter selection and modification improve the method, enhance the applied methodology, and refine the analysis and interpretation of the results. This relationship explains how these parameters ought to be configured, as prior work shows that, when configured, they enable accurate measurement and relevant data collection. Therefore, a comprehensive understanding involves the relationship among the variables; understanding how these parameters function is crucial. We must acknowledge the relationship between the variables. Hence, with appropriate selection and modification of parameters, the research methodology is robust while the analysis and interpretation of the results are thorough.

Table 1. Configuration parameters

Parameter	Nilai
'NUM ARTICLES PREG SPLIT'	5000
'COMPRESSION RATES'	[0.3, 0.5]
'LDA N TOPICS'	5
'MMR LAMBDA'	0.3
'LEXRANK_DAMPING_FACTOR'	0.85
'LEXRANK_THRESHOLD'	0.1
'LEXRANK_MAX_ITER'	100
'LEXRANK_TOL'	1,00E-06
'Weighting Alpha (α)'	0.6

## 3. RESULT

The average ROUGE F1 scores for both approaches across the IndoSUM and Liputan6 datasets are presented in the table. The table offers a comparative analysis of the approaches and their results. The figure highlights these results and adds to the earlier analysis conducted. Further analysis reveals a stark contrast in the efficacy of the two approaches for information summarization across multiple datasets. The algorithm aligns with earlier research on the effectiveness of adequate algorithm selection in textual data processing. Such findings are essential for deepening the understanding of natural language processing and its application in refining information systems for education and publishing [16]. As such, the evidence obtained confirms the

primary hypothesis about the employed methods and expands adaptive user-centric information technology development.

Table 2. Results based on dataset and compression level

Dataset	Tingkat Kompresi	Metode	ROUG E-1	ROUG E-2	ROUG E-L
IndoSum	30%	LexRank	4.092	2.833	3.335
		LexRank+LDA+MMR	<b>4.214</b>	<b>2.916</b>	<b>3.382</b>
	50%	LexRank	3.983	3.022	3.397
		LexRank+LDA+MMR	<b>4.081</b>	<b>3.192</b>	<b>3.505</b>
Liputan6	30%	LexRank	2.868	1.508	2.283
		LexRank+LDA+MMR	<b>3.127</b>	<b>1.694</b>	<b>2.460</b>
	50%	LexRank	2.725	1.530	2.170
		LexRank+LDA+MMR	<b>2.911</b>	<b>1.752</b>	<b>2.346</b>

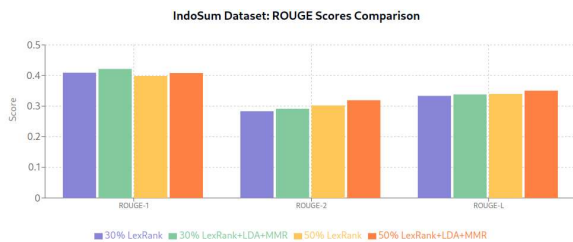


Figure 5. Comparison of rouge f1 average score on indosum dataset

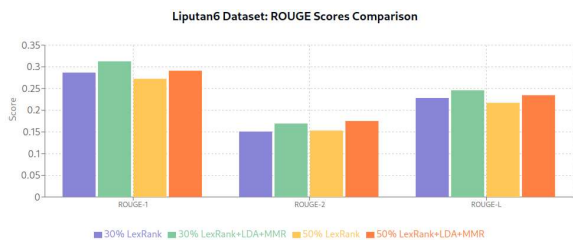


Figure 6. Comparison of rouge f1 average score on liputan6 dataset

#### 4. DISCUSSION

Based on the information in Table 2 and Figures 5 and 6, several points are worth highlighting. The results from the Hybrid Method are nothing short of remarkable. In the two datasets, IndoSum and Liputan6, and at compression levels of 30% and 50%, the integration of LexRank+LDA+MMR outperforms other techniques. Rescue is interdisciplinary, combining policymaking, law, robotics, and computer science. It requires cooperation among international law, humanistic philosophy, and the technologies used in automated rescue operations. This multidisciplinary scope makes Rescue interdisciplinary. The analysis and reasoning use data to develop new insights on the problem of shared citizenship, state-building, and nation-building. The first part concerns the world's hybrid, socio-technical systems. It also basically compares them to non-human-like actors. The technology does not incorporate context reasoning. The second part concentrates on ordinary people who acquire unusual roles and purposes- violence- in their daily life situations. These results generally contribute to researchers and practitioners designing algorithms for more efficient and effective information data systems. The implications of these results are substantial. From the findings illustrated in Table 2 and represented in Figures 5 and 6, the key points are:

Hybrid Method performance: Across both datasets (IndoSum and Liputan6) and at both compression levels (30% and 50%), the combination of LexRank, LDA, and MMR methods consistently outperformed standard LexRank methods on all ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L). The hybrid method suggests that incorporating topic-relevance insights from Landith MMR's diversity mechanism effectively enhances the quality of the generated summaries [17].

Marked Increment in ROUGE-1 and ROUGE-2: The analysis showed the largest improvement in ROUGE-1 and ROUGE-2 scores. Remember that ROUGE-1 measures recall based on word overlap, and ROUGE-2 considers bigram overlap. Elevated scores on these metrics suggest that the summaries generated by the combination approach better capture essential words and significant phrases from the reference summary. May stem from LDA's ability to pinpoint sentences pertinent to the primary topic and MMR's skill in selecting a sequence of sentences that address various facets of the topic without causing excessive redundancy. "Guided by the current investigation, attaining scores of 0.439, 0.183, and 0.406 for ROGUE-1, ROUGE-2, and ROUGE-L, respectively." Considerable progress has been demonstrated with the model achieving all three ROUGE scoring metrics (ROUGE-1, 0.439; ROUGE-2, 0.183; and ROUGE-L, 0.406), confirming the summaries' usefulness. These metrics evaluate the depth of insight and coherence in the summaries processed." Evaluation results suggest that the model has demonstrated its ability to produce coherent and insightful summaries." These additions augment the already powerful claim that the efficiency of the methods employed improves summary quality across domains. Notably, in the IndoSUM dataset, the effectiveness of both methodologies appears to diminish slightly in ROUGE-1 as the compression rate increases from 30% to 50%, even though ROUGE-2 and ROUGE-L remain stable or even improve [18].

Conversely, in the Liputan6 dataset, the traditional LexRank performance declined across all metrics when the compression rate hit 50%. In comparison, with the combined method, the decline observed was smaller, or there was an increase (especially in ROUGE-2). These discrepancies might come from the diverse features of the datasets or the multi-faceted nature of the summarization process at different compression levels. Creating a highly concise summary (30% compression) may be more challenging than a somewhat lengthier summary (50% compression) in preserving all critical information.

**Differences Between Datasets:** Generally, ROUGE scores are higher on the IndoSum datasets than on Liputan6, irrespective of the applied technique. Numerous factors could contribute to this variation. Take, for instance, the summary of the reference to the original article, which is rather personalized, and how other workflows affect the summarization process across different datasets. An associated study showed that data quality and its features substantially affected evaluation results. The integrated method outperformed the traditional LexRank model on both datasets. The algorithm demonstrates that, despite differences across datasets, the general techniques used for summarization are effective. Several other studies support the algorithm by focusing on how to identify and address discrepancies across data sources through normalization [19]. Implications of Using LDA and MMR. These findings align with the theory that integrating a graph centrality measure (LexRank) with a thematic grasp of topic learning methods like LDA, alongside redundancy-aware selection techniques such as MMR, constitutes an efficient strategy for extractive text summarization. In addition to ensuring that the selected sentences focus on the relevant main topic, LDA ensures that the chosen sentences center on the critical main topic. On the other hand, MMR ensures broader subject coverage by selecting diverse, relevant sentences based on subject-specific combination scores. The MMR\_LAMBDA value of 0.3 prioritizes diversity, which appears advantageous in this context, as allocating words to topics relies on the initial value, rendering LDA outcomes generally not entirely reproducible. Even though this combinatory approach improves outcomes, it is crucial to acknowledge that the absolute ROUGE score, particularly in the Liputan dataset 6, still offers opportunities for further improvement to elevate the overall quality of the ranking system. The quality of preprocessing conducted early can influence performance. (For instance, researchers can use filtering techniques to reduce noise and improve the relevance of data.) Although there is improvement with the combination technique, one must remember that the absolute ROUGE score, especially for the Liputan6 dataset, still needs significant work. Assessing preprocessing quality could also affect performance. (such as the application of a superior Bahasa setter or stopword list), parameter choices (number of LDA topics, lambda MMR values, alpha weights), or even the structure of the LDA model itself.

## 5. CONCLUSION

This research comprehensively evaluated the efficacy of integrating LexRank with Latent Dirichlet Allocation (LDA) and Maximal Marginal Relevance (MMR) to summarize extractive

text from Indonesian documents. Through quantitative assessments using ROUGE metrics on the IndoSUM and Liputan6 datasets at varying compression rates of 30% and 50%, researchers determined that the combined approach of LexRank, LDA, and MMR significantly outperformed the conventional LexRank method. The highlighted improvement in performance suggests that considering the relevance of LDA topic modeling and optimizing sentence diversity while maintaining relevance via MMR can significantly improve the extractive summaries resulting from graph-based methods such as LexRank. The summaries generated by this combined approach are generally more informative, as indicated by elevated ROUGE-1 and ROUGE-2 scores, and encompass a broader range of aspects of the topic within the narrative. For upcoming research, there are considerable opportunities to investigate more precise parameter adjustments (e.g., the number of LDA topics, lambda MMR values, and alpha weights), to employ more sophisticated sentence representations (e.g., word embeddings or transformer-based embeddings), and to incorporate additional linguistic features. Moreover, comparing these methodologies with contemporary abstract summarization techniques for Bahasa Indonesia could yield a deeper understanding of the effectiveness and efficiency of emerging techniques [20].

## REFERENCES

- [1] H. K. Pae, "The Impact of Digital Text," in *Script Effects as the Hidden Drive of the Mind, Cognition, and Culture*, vol. 21, in Literacy Studies, vol. 21. , Cham: Springer International Publishing, 2020, pp. 209–217. doi: [10.1007/978-3-030-55152-0\\_11](https://doi.org/10.1007/978-3-030-55152-0_11).
- [2] N. Chatterjee and R. Agarwal, "Studying the Effect of Syntactic Simplification on Text Summarization," *IETE Tech. Rev.*, vol. 40, no. 2, pp. 155–166, Mar. 2023, doi: [10.1080/02564602.2022.2055670](https://doi.org/10.1080/02564602.2022.2055670).
- [3] G. Daga, S. Saha, Y. Shah, and S. J. Nirmala, "Abstractive Text Summarization Using Hybrid Methods," in *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, Kannur, India: IEEE, Aug. 2022, pp. 1294–1300. doi: [10.1109/ICICICT54557.2022.9917994](https://doi.org/10.1109/ICICICT54557.2022.9917994).
- [4] Halimah, Surya Agustian, and Siti Ramadhani, "Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank," *J. CoSciTech Comput. Sci. Inf. Technol.*, vol. 3, no. 3, pp. 371–381, Dec. 2022, doi: [10.37859/coscitech.v3i3.4300](https://doi.org/10.37859/coscitech.v3i3.4300).
- [5] G. Martínez Guzmán, M. B. Bernábe Loranca, C. Cerón Garnica, J. Serrano Pérez, and E. Archundia Sierra, "Application of the LDA Model for Obtaining Topics from the WIKICORPUS," *Comput. Syst.*, vol. 26, no. 1, Mar. 2022, doi: [10.13053/cys-26-1-4171](https://doi.org/10.13053/cys-26-1-4171).
- [6] S. Bellaouar, M. M. Bellaouar, and I. E. Ghada, "Topic Modeling: Comparison of LSA and LDA on Scientific Publications," in *2021 4th International Conference on Data Storage and Data Engineering*, Barcelona Spain: ACM, Feb. 2021, pp. 59–64. doi: [10.1145/3456146.3456156](https://doi.org/10.1145/3456146.3456156).
- [7] N. S. Muningsgar and A. A. Krisnadi, "LexID: The Metadata and Semantic Knowledge Graph Construction of Indonesian Legal Document," *J. Ilmu Komput. Dan Inf.*, vol. 16, no. 1, pp. 15–46, Mar. 2023, doi: [10.21609/jiki.v16i1.1096](https://doi.org/10.21609/jiki.v16i1.1096).

- [8] A. Byerly and T. Kalganova, "Towards an Analytical Definition of Sufficient Data," *SN Comput. Sci.*, vol. 4, no. 2, p. 144, Jan. 2023, doi: [10.1007/s42979-022-01549-4](https://doi.org/10.1007/s42979-022-01549-4).
- [9] P. Pandey, J. Keswani, and S. K. Dash, "Comparative Analysis of Various Techniques Used to Obtain a Suitable Summary of the Document," in *Rising Threats in Expert Applications and Solutions*, vol. 1187, V. S. Rathore, N. Dey, V. Piuri, R. Babo, Z. Polkowski, and J. M. R. S. Tavares, Eds., in *Advances in Intelligent Systems and Computing*, vol. 1187, Singapore: Springer Singapore, 2021, pp. 627–633. doi: [10.1007/978-981-15-6014-9\\_75](https://doi.org/10.1007/978-981-15-6014-9_75).
- [10] R. Van Der Goot, "We Need to Talk About train-dev-test Splits," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 4485–4494. doi: [10.18653/v1/2021.emnlp-main.368](https://doi.org/10.18653/v1/2021.emnlp-main.368).
- [11] R. Wijayanti, M. L. Khodra, and D. H. Widyantoro, "Indonesian Abstractive Summarization using Pre-trained Model," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia: IEEE, Apr. 2021, pp. 79–84. doi: [10.1109/EIConCIT50028.2021.9431880](https://doi.org/10.1109/EIConCIT50028.2021.9431880).
- [12] T. Thieu, H. Do, T. Duong, S. Pu, S. Aakur, and S. Khan, "LexDivPara: A Measure of Paraphrase Quality with Integrated Sentential Lexical Complexity," in *Intelligent Systems and Applications*, vol. 296, K. Arai, Ed., in *Lecture Notes in Networks and Systems*, vol. 296, Cham: Springer International Publishing, 2022, pp. 1–10. doi: [10.1007/978-3-030-82199-9\\_1](https://doi.org/10.1007/978-3-030-82199-9_1).
- [13] D. V. V. Sharmila, B. Natarajan, S. Shalini, and K. C., "An Approach of Statement Compression Using Classifier Algorithm with Improved Efficiency," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India: IEEE, Aug. 2021, pp. 1212–1219. doi: [10.1109/ICESC51422.2021.9532679](https://doi.org/10.1109/ICESC51422.2021.9532679).
- [14] M. Kretinin and G. Nguyen, "Topic Modeling on News Articles using Latent Dirichlet Allocation," in *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)*, Georgioupolis Chania, Greece: IEEE, Aug. 2022, pp. 000249–000254. doi: [10.1109/INES56734.2022.9922609](https://doi.org/10.1109/INES56734.2022.9922609).
- [15] D. Delvin, D. Arisandi, and T. Sutrisno, "Aplikasi Peringkasan Dokumen Menggunakan Metode Maximum Marginal Relevance (MMR)," *J. Ilmu Komput. Dan Sist. Inf.*, vol. 10, no. 1, Mar. 2022, doi: [10.24912/jiksi.v10i1.17820](https://doi.org/10.24912/jiksi.v10i1.17820).
- [16] H. A. Younis, N. I. R. Ruhaiyem, W. Ghaban, N. A. Gazem, and M. Nasser, "A Systematic Literature Review on the Applications of Robots and Natural Language Processing in Education," *Electronics*, vol. 12, no. 13, p. 2864, Jun. 2023, doi: [10.3390/electronics12132864](https://doi.org/10.3390/electronics12132864).
- [17] J. Bishop, Q. Xie, and S. Ananiadou, "GenCompareSum: a hybrid unsupervised summarization method using salience," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 220–240. doi: [10.18653/v1/2022.bionlp-1.22](https://doi.org/10.18653/v1/2022.bionlp-1.22).
- [18] P. Harremoës, "Rate Distortion Theory for Descriptive Statistics," *Entropy*, vol. 25, no. 3, p. 456, Mar. 2023, doi: [10.3390/e25030456](https://doi.org/10.3390/e25030456).
- [19] D. C. Bell, R. C. Boehm, J. Feldhausen, and J. S. Heyne, "A Data Set Comparison Method Using Noise Statistics Applied to VUV Spectrum Match Determinations," *Anal. Chem.*, vol. 94, no. 43, pp. 14861–14868, Nov. 2022, doi: [10.1021/acs.analchem.2c01931](https://doi.org/10.1021/acs.analchem.2c01931).
- [20] M. D. B. Laksana, A. E. Karyawati, L. A. A. R. Putri, I. W. Santiyasa, N. A. Sanjaya Er, and I. G. A. G. A. Kadnyanan, "Text Summarization terhadap Berita Bahasa Indonesia menggunakan Dual Encoding," *JELIKU J. Elektron. Ilmu Komput. Udayana*, vol. 11, no. 2, p. 339, Jul. 2022, doi: [10.24843/JLK.2022.v11.i02.p13](https://doi.org/10.24843/JLK.2022.v11.i02.p13).

## NOMENCLATURES

S	The set of sentences in the document
Si	The i-th sentence in the document
TF IDF(Si)	TF-IDF vector representation of sentence Si
Sim(Si, Sj)	Cosine similarity between sentences Si and Sj
p(u)	Centrality score (LexRank) of sentence u
d	Damping factor in LexRank
N	Total number of sentences in the document
LDA	Latent Dirichlet Allocation
N_topics	Number of topics in the LDA model
Relevance(Si)	Si sentence relevance score (in MMR, using a combination score)
MMR	Maximal Marginal Relevance