Eduvest – Journal of Universal Studies
Volume 5 Number 2, February, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

EVALUATING THE EFFECTIVENESS OF THE LEXRANK AND LSA ALGORITHM IN AUTOMATIC TEXT SUMMARIZATION FOR INDONESIAN LANGUAGE

Galih Wiratmoko
Universitas Muhamadiyah Surakarta, Indonesia
Email: L280220006@student.ums.ac.id

ABSTRACT

The aim of this study is to evaluate how effective the Lexrank algorithm and Latent semantic analysis (LSA) are in automatic text summarization for the Indonesian language. This research focuses on natural language processing and handling of excessive data. We applied both algorithms to generate text summaries using the INDOSUM dataset, which contains about 20,000 news articles in Indonesian with manual summaries. To assess performance, the ROUGE metric was used, which includes aspects of precision, recall, and F1 score. In all tested metrics, LSA outperformed Lexrank. LSA had a precision of 0.57, recall of 0.67, and an F1 score of 0.59, whereas Lexrank had a precision of 0.46, recall of 0.52, and an F1 score of 0.48. These result indicate that LSA is better at gathering important information from the original text than Lexrank.

KEYWORDS *Automatic text summarization, Latent sematic analysis, Lexrank, Bahasa Indonesia*

INTRODUCTION

Understanding large amounts of information quickly is a need that can increase work efficiency. In the mindset of abundant information due to the impact of evolving technology, there is a wealth of online and offline data from various sources that is disseminated daily. There lies a significant challenge in how to present data in a concise and effective manner. Text summarization becomes a solution for how to make large textual information shorter (Khan et al., 2023). Text summarization filters textual information into a concise sentence structure while retaining the message and meaning from its original context. Although manual text summarization can preserve the original meaning of the content, this method requires a relatively long time (Wahab et al., 2023). The solution in the form of automatic text summarization (ATS) has started to gain attention. The importance of ATS in addressing the issue of information overload by providing quick and

How to cite: Galih Wiratmoko. (2025). Evaluating the Effectiveness of the Lexrank and LSA Algorithm in Automatic Text Summarization for Indonesian Language. *Journal Eduvest.* 5(2), 1294-1302
E-ISSN: 2775-3727

efficient summaries is very helpful in fast-paced activities where quick decision-making is crucial (Hernández-Castañeda et al., 2020; Kurniawan & Louvan, 2018).

Automatic text summarization is divided into two types: abstractive and extractive. Extractive text summarization involves selecting sentences or essential information directly from the original document to create a summary. Extractive uses linguistic or statistical features to identify key sentences, while abstractive understands the main concepts of the original text and generates a new summary that captures those concepts in fewer words (Madhuri & Kumar, 2019). ATS are categorized into supervised and unsupervised approaches based on their learning. Supervised ATS algorithms require annotated training data and involve a training phase (Bhuyan et al., 2023; Shah & Desai, 2016; Wang et al., 2023). Unsupervised ATS algorithms, on the other hand, do not require a training phase or training data, thus offering an easier implementation for summarization tasks without needing a labeled dataset (El-Kassas et al., 2021; Kumar et al., 2021).

ATS starts with a text document and extracts or generates summaries using various techniques. Extractive and abstractive summarization are included in these methods. This summarization can be divided into single-document or multi-document summarization. ATS algorithms can also use supervised or unsupervised learning methods. The goal is to produce a concise summary that retains the essential information from the original text, thereby enhancing the efficiency of data retrieval and understanding (Widyassari et al., 2019).

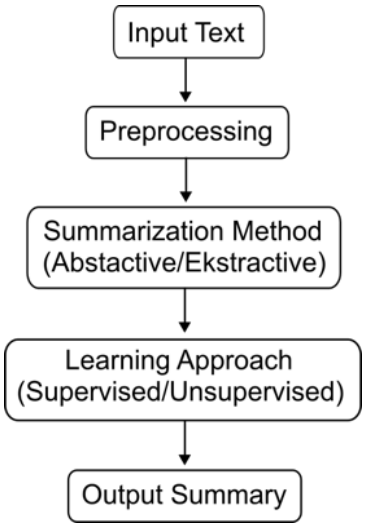

Figure 1. Basic structure of automatic text summarization

The trend in automatic text summarization evolves with the advancement of information. Unsupervised extractive methods such as TF-IDF, LexRank, LSA, and Seq2Seq have been used and produce good summaries (Gunawan et al., 2015; K. Wu et al., 2015). The development of automatic text summarization has evolved using hybrid methods and supervised machine learning, such as BERT, to produce high-quality summaries by reducing redundancy (Fan et al., 2023). Recent advancements include the use of Generative Pre-trained Transformer (GPT) models

that provide accurate and contextual summaries, as proven through evaluations using ROUGE against T5 and GPT models (Dhivyaa et al., 2022).

RESEARCH METHOD

This study examines a comparative analysis between LexRank and Latent Semantic Analysis (LSA). The aim is to assess their ability to produce concise and consistent text summaries. Our analysis is structured in several stages, including data loading, pre-processing, summarization, and evaluation, to achieve the desired standard summaries (G.-H. Wu & Guo, 2015).

Dataset

The INDOSUM dataset, which includes about 20,000 Indonesian-language news articles with manual summaries organized in various categories, was chosen for this study due to its broad representation. The purpose of this dataset is to enhance research on natural language processing in the Indonesian language.

Pre-processing

The data used for this study was collected through a thorough pre-processing process, which includes text cleaning, tokenization, and stopword removal. Stemming. This process ensures that the data processed by the text summarization algorithms is structured and clean.

Table 1. Pre-Processing on the Dataset

Article before text cleaning, tokenization, stopword removal, and stemming.
Jakarta, CNN Indonesia - - Dokter Ryan Thamrin, yang terkenal lewat acara Dokter Oz Indonesia, meninggal dunia pada Jumat (4/8) dini hari. Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan menyebut kawannya itu sudah sakit sejak setahun yang lalu. Lula menuturkan, sakit itu membuat Ryan mesti vakum dari semua kegiatannya, termasuk menjadi pembawa acara Dokter Oz Indonesia. Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru, Riau untuk menjalani istirahat. ‘Setahu saya dia orangnya sehat, tapi tahun lalu saya dengar dia sakit.
Article after text cleaning, tokenization, stopword removal, and stemming.
“Jakarta, CNN Indonesia – Dokter Ryan Thamrin, kenal acara Dokter Oz Indonesia, tinggal dunia Jumat (4/8) dini. Dokter Lula Kamal, selebriti kerja Ryan, sebut kawan sakit tahun lalu. Lula, sakit buat Ryan vakum giat, bawa acara Dokter Oz Indonesia. Kondisi buat Ryan kembali kampung halaman Pekanbaru, Riau, jalani istirahat. ‘Tau saya orang sehat, tahun dengar sakit.’

Lexrank Summarization

The LexRank algorithm is an unsupervised approach to automatic text summarization based on graph theory. At its core, LexRank determines the importance of each sentence in the text to identify and extract the most informative sentences for the summary. The process begins by treating each sentence in the document as a node in a graph. The edges between these nodes represent the similarity between

sentences, often calculated using the cosine similarity measure of the TF-IDF (Term Frequency-Inverse Document Frequency) representations of the sentences (G.-H. Wu & Guo, 2015).

This study investigates the domain of automatic text summarization using the Lexrank algorithm to amalgamate key elements from articles into a concise summary. It begins by loading data, where the system reads selected jsonl files and loads the first three articles to ensure thorough analysis. This method then progresses in a structured manner for each article. First, content is gathered from paragraphs, forming a consistent body of text. The next critical step is the removal of stop words. This occurs when the text is encoded into sentences and words, and stop words are removed from each sentence to enhance the text's relevance and clarity.

1) Start : Begin text summarization
2) Load data : Read dataset file
3) For each article :
    - Extract content : Compile text content from article paragraphs
    - Stop word removal : tokenize content into sentences and words, remove stop words from each sentence, then construct text.
    - Summarization with Lexrank : Convert the processed text into a PlaintextParser object, Initialize Lexrank summarizer, Determine the number of sentences to be included in the summary, Generate summary by selecting top sentences according to Lexrank.
    - Extract Gold summary : construct the gold summary from the dataset
    - Output : Print original content, print lexrank-generated summary, Print gold summary
    - End for each article
4) Complete: Finish summarization.

The summarization process includes the application of LexRank, an advanced algorithm specifically created for the task of summarization. It involves activating the LexRank Summarizer, converting the processed text into a PlaintextParser object, and determining the number of sentences to be included in the summary. This criteria selects a summary that is concise and relevant with a minimum of five sentences or the total number of existing sentences. The summarization culminates with the generation of a summary that selects the best sentences based on LexRank, which captures the main topics of the article.

Table 2. Summary Generated by the Lexrank Model

Artikel
Jakarta , CNN Indonesia - - Dokter Ryan Thamrin , yang terkenal lewat acara Dokter Oz Indonesia , meninggal dunia pada Jumat (4 / 8) dini hari . Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan menyebut kawannya itu sudah sakit sejak setahun yang lalu . Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia . Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru , Riau untuk menjalani istirahat . " Setahu saya dia orangnya sehat , tapi tahun lalu saya dengar dia sakit……..

LexRank model summary
Dokter Ryan Thamrin , terkenal acara Dokter Oz Indonesia , meninggal dunia Jumat (4 / 8) . Lula , sakit Ryan mesti vakum kegiatannya , pembawa acara Dokter Oz Indonesia . (Karena) sakitnya , langsung pulang Pekanbaru , jenguk susah . Barangkali istirahat , ya , Jakarta susah isirahatnya, Lula CNNIndonesia.com , Jumat (4 / 8) . Saya , barangkali penyakit berbeda , penyebab kematiannya beda penyakit.
Gold summary
Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan Thamrin menyebut kawannya itu sudah sakit sejak setahun yang lalu . Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia . Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru , Riau untuk menjalani istirahat.

The process involves extracting the gold summary from the article data, which allows for a comparison between the algorithm-generated summary and the ideal summary provided in the data. The output phase of the process involves printing the original content, the LexRank-generated summary, and the gold summary, providing a comprehensive picture of the summarization's effectiveness.

Latent sematic analysis summarization

The steps used in the process of automatic text summarization using LSA are not much different. The process begins with loading the first three articles from the .jsonl file. Afterward, content is extracted and simplified by removing stop words. Next, the text is processed using LSA to identify the key sentences to be used in the summary. To determine how effective it is, the last step is to compare the created summary with the gold summary. The aim of the entire procedure is to produce a summary that is concise and useful, which aids the field of natural language processing.

Table 3. Summary Generated by Lsa

Articel
Jakarta , CNN Indonesia - - Dokter Ryan Thamrin , yang terkenal lewat acara Dokter Oz Indonesia , meninggal dunia pada Jumat (4 / 8) dini hari . Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan menyebut kawannya itu sudah sakit sejak setahun yang lalu . Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia . Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru , Riau untuk menjalani istirahat . " Setahu saya dia orangnya sehat , tapi tahun lalu saya dengar dia sakit……..
LSA Summary
Dokter Ryan Thamrin , yang terkenal lewat acara Dokter Oz Indonesia , meninggal dunia pada Jumat (4 / 8) dini hari . Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia . (Karena) sakitnya , ia langsung pulang

ke Pekanbaru , jadi kami yang mau jenguk juga susah . Barangkali mau istirahat , ya betul juga , kalau di Jakarta susah isirahatnya , " kata Lula kepada CNNIndonesia.com , Jumat (4 / 8) . Ryan Thamrin terkenal sebagai dokter yang rutin membagikan tips dan informai kesehatan lewat tayangan Dokter Oz Indonesia.
Gold Summary
Dokter Lula Kamal yang merupakan selebriti sekaligus rekan kerja Ryan Thamrin menyebut kawannya itu sudah sakit sejak setahun yang lalu . Lula menuturkan , sakit itu membuat Ryan mesti vakum dari semua kegiatannya , termasuk menjadi pembawa acara Dokter Oz Indonesia . Kondisi itu membuat Ryan harus kembali ke kampung halamannya di Pekanbaru , Riau untuk menjalani istirahat.

Model Evaluation

To evaluate the performance of the text summarization system, this study uses NLP metrics such as ROUGE (Recall-Oriented Understudy for GISTING Evaluation), which compares the similarity between the summaries generated by the model and the original summaries. The detailed steps in the ROUGE calculation are as follows:

- Calculating LCS(Longest common subsequence) : The LCS between two summaries, the human summary and the model summary, is calculated. The LCS is the longest sequence of elements in both summaries in the same order, although not necessarily consecutive. This provides a measure of similarity based on content and sequence.
- Calculating Precision : Precision is determined as the ratio of the length of the summary generated by the model to the length of the LCS. It indicates the proportion of information in the model summary that is relevant to the human summary.

$$\text{Precision} = \frac{LCS}{length\ of\ the\ model\ summary}$$

- Calculating Recall : Recall is calculated by dividing the length of the LCS by the length of the human summary. It is a way to determine how completely the information in the human summary is represented in the model summary.

$$\text{Recall} = \frac{LCS}{Length\ of\ the\ human\ summary}$$

- Calculating F1-Score : The F1-Score is the harmonic mean of precision and recall, which provides a measure that balances both elements. It is used to measure overall how well the model summary captures important data from the human summary.
-

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

RESULT AND DISCUSSION

Using the INDOSUM dataset, here are the results of LexRank and LSA summarization scores conducted with various scenarios. In initial testing, ROUGE was used to calculate Precision, Recall, and F1 scores tested on the first 5 articles, concluding with two models, namely LexRank and LSA.

Table 4. Comparasion of Summary Result Sample

Articel	Lexrank			LSA		
	p	r	f	p	r	f
A1	0.50	0.42	0.46	0.43	0.69	0.53
A2	0.42	0.41	0.41	0.33	0.53	0.41
A3	0.50	0.70	0.58	0.77	0.77	0.59
A748	0.64	0.71	0.67	0.83	0.83	0.84
A749	0.35	0.52	0.42	0.68	0.68	0.65
A750	0.45	0.60	0.51	0.41	0.57	0.48
Average	**0.47**	**0.56**	**0.50**	**0.57**	**0.67**	**0.58**

The comparison table shows that the Latent Semantic Analysis (LSA) method appears to have a higher recall score compared to Lexrank. This suggests that LSA may be better at gathering essential information from the original summary but might also experience a decrease in precision, indicating the presence of additional irrelevant information. On the other hand, Lexrank demonstrates a more balanced distribution between precision and recall.

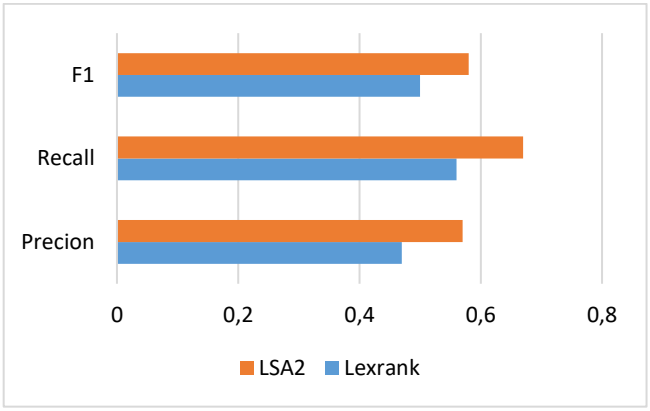

Figure 2. ROUGE result from sample article

In the next stage, testing was conducted on the INDOSUM testing dataset, which comprises 3750 articles, to compare two popular automatic text summarization models, LexRank and Latent Semantic Analysis (LSA). The evaluation was performed using the metrics of precision, recall, and F-measure. The LexRank model showed a precision of 0.46, a recall of 0.52, and an F-measure of 0.48.

Meanwhile, the LSA model demonstrated improved performance with a precision of 0.57, a recall of 0.67, and an F-measure of 0.59.

Table 5. Final Comparison Result of Lexrank And LSA Models

	Lexrank			LSA		
	p	**r**	**f**	**p**	**r**	**f**
Result	0.46	0.52	0.48	0.57	0.67	0.59

The evaluation results indicate that the Latent Semantic Analysis (LSA) model performs better than LexRank in terms of all the metrics used. Notably, LSA excels with a significant margin in recall, indicating that this model is more effective in capturing the important sentences that should be included in the summary. Although both models have room for improvement, especially in increasing precision to select fewer irrelevant sentences, this data suggests that LSA is a more recommended choice for automatic text summarization on the dataset used in this study.

CONCLUSION

This study evaluated the effectiveness of two unsupervised algorithms—LexRank and Latent Semantic Analysis (LSA)—in automatic text summarization for the Indonesian language. Using the INDOSUM dataset, the results showed that LSA outperformed LexRank across all metrics, including precision, recall, and F1 score. Specifically, LSA demonstrated a higher recall, indicating its superior ability to capture essential information from the original text. While both models showed room for improvement, especially in terms of precision, the study recommends LSA as a more effective choice for automatic text summarization tasks, particularly when dealing with large-scale datasets in the Indonesian language. The findings highlight the importance of refining these algorithms for better performance in handling complex natural language processing tasks.

REFERENCES

Bhuyan, S. S., Mahanta, S. K., Pakray, P., & Favre, B. (2023). Textual entailment as an evaluation metric for abstractive text summarization. *Natural Language Processing Journal*, *4*, 100028.

Dhivyaa, C. R., Nithya, K., Janani, T., Kumar, K. S., & Prashanth, N. (2022). Transliteration based generative pre-trained transformer 2 model for Tamil text summarization. *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679.

Fan, J., Tian, X., Lv, C., Zhang, S., Wang, Y., & Zhang, J. (2023). Extractive social media text summarization based on MFMMR-BertSum. *Array*, *20*, 100322.

Gunawan, F. E., Juandi, A. V., & Soewito, B. (2015). An automatic text summarization using text features and singular value decomposition for

popular articles in Indonesia language. *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 27–32.

Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., & Millán-Hernández, C. E. (2020). Extractive automatic text summarization based on lexical-semantic keywords. *IEEE Access*, *8*, 49896–49907.

Khan, B., Shah, Z. A., Usman, M., Khan, I., & Niazi, B. (2023). Exploring the landscape of automatic text summarization: a comprehensive survey. *IEEE Access*.

Kumar, Y., Kaur, K., & Kaur, S. (2021). Study of automatic text summarization approaches in different languages. *Artificial Intelligence Review*, *54*(8), 5897–5929.

Kurniawan, K., & Louvan, S. (2018). Indosum: A new benchmark dataset for indonesian text summarization. *2018 International Conference on Asian Language Processing (IALP)*, 215–220.

Madhuri, J. N., & Kumar, R. G. (2019). Extractive text summarization using sentence ranking. *2019 International Conference on Data Science and Communication (IconDSC)*, 1–3.

Shah, P., & Desai, N. P. (2016). A survey of automatic text summarization techniques for Indian and foreign languages. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 4598–4601.

Wahab, M. H. H., Ali, N. H., Hamid, N. A. W. A., Subramaniam, S. K., Latip, R., & Othman, M. (2023). A review on optimization-based automatic text summarization approach. *IEEE Access*, *12*, 4892–4909.

Wang, M., Xie, P., Du, Y., & Hu, X. (2023). T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions. *Applied Sciences*, *13*(12), 7111.

Widyassari, A. P., Affandy, A., Noersasongko, E., Fanani, A. Z., Syukur, A., & Basuki, R. S. (2019). Literature review of automatic text summarization: research trend, dataset and method. *2019 International Conference on Information and Communications Technology (ICOIACT)*, 491–496.

Wu, G.-H., & Guo, Y.-T. (2015). An enhanced LSA-based approach for update summarization. *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 493–497.

Wu, K., Shi, P., & Pan, D. (2015). An approach to automatic summarization for chinese text based on the combination of spectral clustering and LexRank. *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 1350–1354.