Syntax Admiration: p-ISSN 2722-7782 | e-ISSN 2722-5356 Vol. 5, No. 9, September 2024

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.

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, Shah, Usman, Khan, & Niazi, 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 efficient summaries is very helpful in fast-paced activities where quick decision-making is crucial.

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 [3]. ATS are categorized into supervised and unsupervised approaches based on their learning. Supervised ATS algorithms require annotated training data and involve a training phase. 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.

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).

In the early stages of research on natural language processing in automatic text summarization, the focus was generally on algorithms that assess each sentence in the text based on statistics such as the occurrence of words in a sentence [9]. Automatic text summarization is a technique for generating shorter text from longer text that contains important information (Shah & Desai, 2016).

Dataset, In automatic text summarization, the dataset is a collection of data used to create the summarization system and assess its performance (Gunawan, Juandi, & Soewito, 2015). Although most algorithms are designed for English, Indonesian has specific issues. Document Understanding Conference (DUC), 2021-2007, Text Analysis Conference (TAC), Opinosis, CAST, CNN Corpus Dataset, Gigaword 5, and CNN/Daily Mail Corpus are some examples of popular datasets (Bhuyan, Mahanta, Pakray, & Favre, 2023).

Although Indonesian is one of the most widely spoken languages in the world, there are few datasets for natural language processing tasks. INDOSUM, which contains about 20,000 news articles with manual summaries, is the most commonly used dataset for automatic text summarization. The aim of INDOSUM is to encourage natural language (NLP) research in Indonesia and assist in the development of more advanced natural language processing methods (Kurniawan & Louvan, 2018).

PreProcessing, Text pre-processing is a crucial process that enables summarization algorithms to analyze raw data in a more structured format. To achieve accurate and efficient summarization, this process enhances the quality of data to be summarized. Pre-processing includes data cleaning, tokenization, stop word removal, lemmatization and stemming processes, and splitting the text into sentences.

Summarization method, To create a summary, extractive text summarization methods use important sentences from the source text. This method is simpler because it only involves selecting significant parts of the original document, maintaining grammatical correctness and a high level of accuracy. Studies have been conducted on extractive summarization for Indonesian-language texts, and the results vary (Hernández-Castañeda, García-Hernández, Ledeneva, & Millán-Hernández, 2020). Abstract models focus on interpreting documents comprehensively and conveying the document's content

in altered sentences. This not only simplifies information but also conveys concepts or ideas in a different sentence structure. The more difficult process is one that requires special abilities such as rearranging and generalizing data (Mridha et al., 2021).

Evaluation, In the field of natural language processing, several key metrics used to evaluate the performance of automatic text summarization systems include ROUGE, BLEU, METEOR, and CIDEr. Specifically, ROUGE evaluation, or Recall-Oriented Understudy for Gisting Evaluation, is conducted by comparing how similar the summaries generated by the system are to the reference summaries. The most commonly used versions of ROUGE include ROUGE-N, which calculates the n-gram overlap between summaries; ROUGE-L, which assesses structural consistency based on the longest common subsequence between two summaries; and ROUGE-SU, which incorporates both skip-bigrams and unigrams, allowing for a more diverse and flexible assessment (Ay, Ertam, Fidan, & Aydin, 2023).

Latent Semantic Analysis (LSA) is a computational technique used to understand the essence of words and phrases through statistical analysis of large amounts of text data. This method is used for various purposes, such as classifying texts and finding keywords (Wu, Shi, & Pan, 2015). Gong and Liu first used LSA in document summarization tasks. Subsequently, Steinberger developed this method for the task of update summarization. Steinberger's main method involves identifying hidden themes in a collection of documents, assessing the novelty of emerging topics by comparing them to previous topics, and selecting sentences that represent the most recent and important topics (Dhivyaa, Nithya, Janani, Kumar, & Prashanth, 2022).

Lexrank is an automatic summarization method based on the PageRank algorithm. It assesses the importance of sentences in the text using a weighted cosine similarity graph of sentences. This method states that not only the number of connected sentences matters but also the importance or significance of the connected sentences (Fan et al., 2023). Using a stochastic matrix derived from the cosine similarity between sentences, Lexrank treats each sentence as a node in a graph with similarity scores on its edges. To calculate the LexRank score, an iterative process is used, where the score of each sentence is evenly distributed among other sentences based on their similarity

Method Research

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. 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. 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.

Tabel 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.'

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.

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.

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.

Tabel 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.

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.

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.

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

Resulth 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 3. 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.

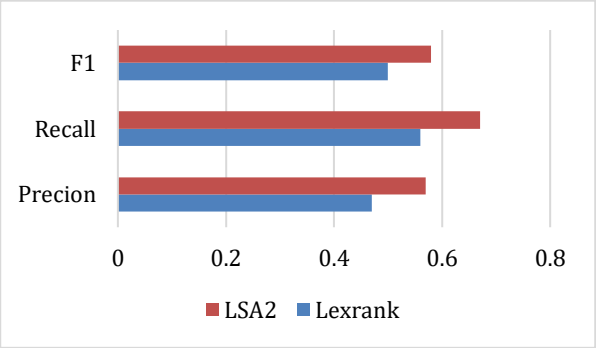

Figure 1. ROUGE Result from Sample Article

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.

Tabel 4. 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

Conclusion

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.

BIBLIOGRAFI

Ay, Betul, Ertam, Fatih, Fidan, Guven, & Aydin, Galip. (2023). Turkish abstractive text document summarization using text to text transfer transformer. *Alexandria Engineering Journal*, *68*, 1–13. https://doi.org/10.1016/j.aej.2023.01.008.

Bhuyan, Swagat Shubham, Mahanta, Saranga Kingkor, Pakray, Partha, & Favre, Benoit.

(2023). Textual entailment as an evaluation metric for abstractive text summarization. *Natural Language Processing Journal*, *4*, 100028. https://doi.org/10.1016/j.nlp.2023.100028.

Dhivyaa, C. R., Nithya, K., Janani, T., Kumar, K. Sathis, & 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. https://doi.org/10.1109/ICCCI54379.2022.9740991

Fan, Junqing, Tian, Xiaorong, Lv, Chengyao, Zhang, Simin, Wang, Yuewei, & Zhang, Junfeng. (2023). Extractive social media text summarization based on MFMMR-BertSum. *Array*, *20*, 100322. https://doi.org/10.1016/j.array.2023.100322.

Gunawan, Fergyanto E., Juandi, Adrian Victor, & Soewito, Benfano. (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. https://doi.org/10.1109/ISITIA.2015.7219948.

A. N. Enhanced, L. A. For, and U. Summarization, "An enhanced lsa-based approach for update summarization," pp. 493–497.

Hernández-Castañeda, Ángel, García-Hernández, René Arnulfo, Ledeneva, Yulia, & Millán-Hernández, Christian Eduardo. (2020). Extractive automatic text summarization based on lexical-semantic keywords. *IEEE Access*, *8*, 49896–49907.

J. N. Madhuri, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf. Data Sci. Commun., pp. 1–3, 2019.

Khan, Bilal, Shah, Zohaib Ali, Usman, Muhammad, Khan, Inayat, & Niazi, Badam. (2023). Exploring the landscape of automatic text summarization: a comprehensive survey. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3322188

Kurniawan, Kemal, & Louvan, Samuel. (2018). Indosum: A new benchmark dataset for indonesian text summarization. *2018 International Conference on Asian Language Processing (IALP)*, 215–220. https://doi.org/10.1109/IALP.2018.8629109.

Mridha, Muhammad Firoz, Lima, Aklima Akter, Nur, Kamruddin, Das, Sujoy Chandra, Hasan, Mahmud, & Kabir, Muhammad Mohsin. (2021). A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, *9*, 156043–156070. https://doi.org/10.1109/ACCESS.2021.3129786.

Shah, Prachi, & Desai, Nikita 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. https://doi.org/10.1109/ICEEOT.2016.7755587.

Y. Kumar, K. Kaur, and S. Kaur, Study of automatic text summarization approaches in different languages, vol. 54, no. 8. Springer Netherlands, 2021. doi: 10.1007/s10462-021-09964-4.

Wahab, Muhammad Hafizul H., Ali, Nor Hafiza, Hamid, Nor Asilah Wati A., Subramaniam, Shamala K., Latip, Rohaya, & Othman, Mohamed. (2023). A Review on Optimization-Based Automatic Text Summarization Approach. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3348075.

Widyassari, Adhika Pramita, Affandy, Affandy, Noersasongko, Edy, Fanani, Ahmad Zainul, Syukur, Abdul, & Basuki, Ruri Suko. (2019). Literature review of automatic text summarization: research trend, dataset and method. *2019 International Conference on Information and Communications Technology (ICOIACT)*, 491–496. https://doi.org/10.1109/ICOIACT46704.2019.8938454.

W. S. El-kassas, C. Salama, A. Rafea, and H. K. Mohamed, "Automatic Text Summarization : A Comprehensive Survey," no. July, 2020, doi: 10.1016/j.eswa.2020.113679.

Wu, Kang, Shi, Ping, & Pan, Da. (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. https://doi.org/10.1109/FSKD.2015.7382140