the 15th International Conference on Green Technology
Faculty of Science & Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia
29 – 30 July, 2025

Application of the TextRank Algorithm for Indonesian News Text Summarization Using Word2Vec and LSA

Prana Wijaya Pratama Nandana[1]*, Muhammad Faisal[2]
[12]Departement of Informatics Engineering, UIN Maulana Malik Ibrahim Malang
210605110120@student.uin-malang.ac.id

***Abstract*- In the digital era, the increasing volume of online news makes it difficult for readers to identify relevant and trustworthy information. This study proposes an extractive summarization system for Indonesian news articles by combining the TextRank algorithm with Word2Vec and Latent Semantic Analysis (LSA). Unlike previous methods such as LexRank, YAKE, and Genetic Algorithm (GA), which have shown limited F1-scores around 0.453, this study introduces a novel integration of semantic and graph-based techniques to improve summary relevance and coherence. Experiments were conducted using the IndoSum dataset comprising 5,000 news articles. The system was evaluated using ROUGE metrics and manual validation. The best performance was achieved at a 30% compression level, yielding ROUGE-1 of 0.4808, ROUGE-2 of 0.3433, and ROUGE-L of 0.4675. Manual evaluation also confirmed that the generated summaries were more informative and readable compared to those from other compression levels. These results indicate that the proposed approach contributes to the advancement of automatic summarization techniques in the field of natural language processing, particularly for the Indonesian language.**

Keywords*: *Extractive, Latent Semantic Analysis, Summarization, TextRank, Word2Vec

I. INTRODUCTION

Online news refers to the latest reports on various events published through digital platforms, accessible anytime and anywhere as long as an internet connection is available [1]. According to the Reuters Institute survey conducted in 2023, online news has become one of the primary sources of information for the Indonesian public, with a consumption rate reaching 84% [2]. However, the large volume of news published daily poses a new challenge, namely the difficulty for readers to filter relevant and credible news amidst the rampant phenomenon of fake news, which can degrade the quality of information [3]. Along with the rapid growth of information, the ability to convey the essence of news quickly has become increasingly important [4].

One solution to address this issue is automatic text summarization technology, which is capable of presenting the core information in a concise and efficient manner, allowing readers to obtain important information rapidly [5]. Various studies have been conducted to develop this technology. One such study was conducted by Wijaya and Girsang [6], who proposed a hybrid method combining the LexRank and YAKE algorithms for summarizing Indonesian news articles. This method was tested on the IndoSum dataset containing 5000 news articles and yielded an F1-score of 0.453, outperforming the individual use of either LexRank or TextRank + GA methods.

Although the study showed improved performance, the resulting F1-score still indicates potential for further enhancement. This suggests that existing summarization systems have yet to fully capture the essence of information with a high level of accuracy. Therefore, a more in-depth approach is required to understand the context and inter-sentence relationships, so that the generated summaries become more accurate and relevant.

Improving summary quality is not only technically important but also has significant social implications. In modern society, information spreads rapidly and often without adequate verification processes [7]. Therefore, the ability to present summaries that are concise, clear, and accurate is crucial, not only for efficiency but also to support the dissemination of correct and trustworthy information.

In response to this need, the present study develops an automatic summarization system for Indonesian-language news by integrating the TextRank algorithm, Word2Vec, and Latent Semantic Analysis (LSA). Unlike previous studies that typically combined only two methods, this approach integrates three techniques simultaneously: graph-based sentence ranking, semantic representation, and dimensionality reduction. This study also employs the IndoSum dataset, consisting of 5,000 articles the same quantity used in prior research. However, to date, no study has systematically applied the combination of these three methods to this dataset. This indicates a research gap that the proposed approach seeks to address

II. RESEARCH METHOD

The automatic text summarization system developed in this study comprises several main stages, as illustrated in Figure. 1. The process begins with data input, followed by text preprocessing, feature extraction using Word2Vec and Latent Semantic Analysis (LSA), calculation of sentence similarity using cosine similarity, application of the TextRank algorithm for selecting important sentences, and concludes with both automated and manual evaluation of the results.

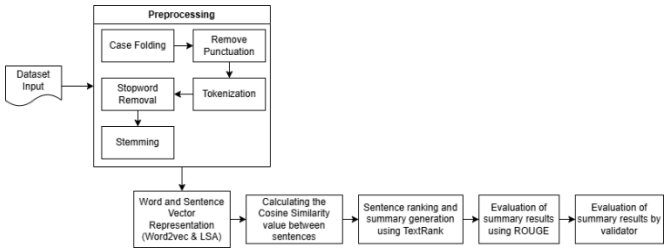

Fig. 1. System Design

A. Dataset

This study utilizes a total of 5,000 Indonesian-language news articles obtained from the IndoSum dataset, covering various topics such as politics, economics, health, technology, sports, and culture [8]. The IndoSum dataset is accompanied by manually written summaries created by native Indonesian speakers, which are used as reference for evaluation purposes.

B. Text Processing

This stage aims to clean the text to make it ready for analysis. The preprocessing steps include case folding, punctuation removal, tokenization, stopword removal, and stemming. The main objective is to produce text that is more structured and consistent, making it suitable for the subsequent feature extraction stage.

C. Word2Vec

Word2Vec is utilized to convert words into high-dimensional vector representations that reflect their meanings based on context [9]. In this study, the Skip-gram model is trained using the following parameters: vector_size = 300, window = 6, min_count = 2, negative = 15, learning_rate = 0.03, over 100 epochs with 4 worker threads. The use of a window size of 6 and a vector dimension of 300 refers to a commonly adopted configuration in word embedding research, as it is considered to provide a balanced trade-off between performance and computational efficiency [10]. After training is completed, the resulting word vectors are averaged to construct the vector representation of each sentence.

D. Latent Semantic Analysis

After obtaining the sentence vectors, dimensionality reduction is performed using Latent Semantic Analysis (LSA) to simplify the semantic representation and reduce noise. The sentence vector matrix is reduced using the Singular Value Decomposition (SVD) technique to 180 principal components [11]. The reduced representations are then used to calculate inter-sentence similarity.

E. Cosine Similarity

The similarity between sentences is measured using cosine similarity applied to the reduced sentence vectors [12]. These similarity scores are used as edge weights in the TextRank graph. Cosine similarity is calculated using the following formula (1)[13]:

$$cosine\_similarity = \frac{A \cdot B}{norm\_a \times norm\_b} \quad (1)$$

Where A and B are vectors representing two different sentences, A · B denotes the dot product of the vectors, and ‖A‖ and ‖B‖ are the norms (magnitudes) of the respective vectors, calculated using the Euclidean formula:

$$norm\_a = \sqrt{\sum_{i=1}^{n} a_i^2} \quad (2)$$

$$norm\_b = \sqrt{\sum_{i=1}^{n} b_i^2} \quad (3)$$

The results of the cosine similarity calculation are used to determine the edge weights in the TextRank algorithm graph. The higher the cosine similarity between two sentences, the stronger the semantic relationship between them, which increases the likelihood that the sentence will be included in the summary.

F. TextRank

TextRank constructs a weighted graph in which each node represents a sentence, and the edges between nodes are weighted based on cosine similarity values. TextRank scores are computed iteratively until convergence, and the highest-scoring sentences are then selected as the summary[14]. In this study, a damping factor of 0.85 and a maximum of 500 iterations are used.

The computation of TextRank scores is defined by the following formula (4) [15]:

$$S(V_i) = (1 - \mathrm{d}) + \mathrm{d} \cdot \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_j \in In(V_i)} w_{jk}} S(V_j) \quad (4)$$

Where $S(V_i)$ is the TextRank score of node V_J d is the damping factor (typically set to 0.85), w_{ji} is the edge weight between nodes V_J and V_ii, and $In(V_i)$ and $Out(V_j)$ are the sets of nodes pointing to and from a given node, respectively. The iteration continues until the scores converge or the maximum number of iterations is reached. Sentences with the highest scores are selected according to the desired summary compression level to form a final summary that is both informative and coherent. With this approach, TextRank is capable of generating summaries without requiring training data, making it effective for Indonesian news texts, which typically exhibit clear structure and high information density.

G. Evaluation

Evaluation is conducted using two approaches—automatic and manual to ensure the quality of the generated summaries. The automatic evaluation employs the ROUGE (Recall-

Oriented Understudy for Gisting Evaluation) metric, which measures n-gram overlap between system-generated summaries and reference summaries based on ROUGE-1, ROUGE-2, and ROUGE-L[16]. The evaluation metrics include precision, recall, and F1-score as performance indicators of the system [17].

In addition to automatic evaluation, manual evaluation is conducted to assess aspects that may not be fully captured by automated metrics. Manual assessment is performed by a validator on 370 news samples, randomly selected from the total of 5,000 articles using the Simple Random Sampling method to ensure objective and representative results [18]. The sample size is determined using the Slovin formula with a 5% margin of error, resulting in a total of 370 data points. The validator then evaluates the quality of the summaries based on these samples. The Slovin formula used for determining the sample size is presented in Equation (5) [19].

$$n = \frac{N}{1 + Ne^2} \quad (5)$$

III. RESULTS AND DISCUSSION

A. Summary Result

The system is evaluated using 5,000 news articles from the IndoSum dataset. The testing is conducted at three summary compression levels: 10%, 20%, and 30% of the original text length. Table I presents a comparison between the original text lengths and the corresponding summary outputs.

TABLE I. DIFFERENCES IN NUMBER OF WORDS AND SENTENCES

Scenario	Number of Word	Number of Sentence
Original News Text	344	20
System summary (10%)	39	3
System summary (20%)	81	4
System summary (30%)	120	6

TABLE II. EXAMPLE OF SUMMARY RESULT

Text Type	Summary Result
Original Text	Jakarta , CNN Indonesia - - Timnas Argentina kembali gagal meraih kemenangan tanpa kehadiran Lionel Messi . Kali ini tim Tango ditahan imbang Peru 2 - 2 di Stadion Nasional , Lima , Kamis (6 / 10) malam waktu setempat atau Jumat pagi WIB . Argentina membuang keunggulan dua kali saat menghadapi Peru . Ramiro Funes Mori membawa Argentina unggul lewat gol pada menit ke - 16 . Bek Everton itu memanfaatkan kemelut di kotak penalti Peru lewat skema sepak pojok . Keunggulan satu gol Argentina bertahan hingga jeda babak pertama . Di babak kedua , tepatnya menit ke - 58 , tuan rumah berhasil menyamakan kedudukan lewat kapten tim Jose Paolo Guerrero . Usai menerima umpan Miguel Trauco , Guerrero berhasil menahan bola dengan dada dan lolos dari kawalan Funes Mori . Mantan penyerang Bayern Munich itu kemudian melepaskan tendangan mendatar yang tidak bisa dihentikan kiper Sergio Romero . Argentina kembali unggul lewat Gonzalo Higuain pada menit ke - 78 . El Pipita berhasil meneruskan umpan terobosan Pablo Zabaleta dan mencungkil bola melewati kiper Pedro Gallese . Sayang , Argentina yang ditangani Edgardo Bauza tidak mampu mempertahankan keunggulan . Pada menit ke - 84 , Peru mendapat tendangan penalti yang dieksekusi dengan sempurna oleh Christian Cueva setelah Mori menjatuhkan Guerrero . Hasil imbang 2 - 2 membuat Argentina gagal naik dari peringkat lima klasemen sementara kualifikasi Piala Dunia 2018 zona Conmebol . Argentina kini mengoleksi 16 poin , tertinggal tiga poin dari Uruguay di puncak klasemen . Argentina untuk kali kedua beruntun meraih hasil imbang . Sebelumnya , Argentina ditahan imbang Venezuela 2 - 2 , 6 September lalu . Menariknya , di kedua laga tersebut Messi tidak tampil karena cedera . Messi saat ini sedang menjalani pemulihan cedera pangkal paha selama tiga pekan yang didapatnya saat melawan Atletico Madrid . Penyerang Barcelona itu juga dipastikan absen ketika Argentina menghadapi Paraguay di Cordoba , 11 Oktober mendatang .(har)
Summary System (10%)	Keunggulan satu gol Argentina bertahan hingga jeda babak pertama . Ramiro Funes Mori membawa Argentina unggul lewat gol pada menit ke – 16.
System summary (20%)	Keunggulan satu gol Argentina bertahan hingga jeda babak pertama . Ramiro Funes Mori membawa Argentina unggul lewat gol pada menit ke - 16 . Sayang , Argentina yang ditangani Edgardo Bauza tidak mampu mempertahankan keunggulan . Jakarta , CNN Indonesia - - Timnas Argentina kembali gagal meraih kemenangan tanpa kehadiran Lionel Messi.
System summary (30%)	Keunggulan satu gol Argentina bertahan hingga jeda babak pertama . Ramiro Funes Mori membawa Argentina unggul lewat gol pada menit ke - 16 . Sayang , Argentina yang ditangani Edgardo Bauza tidak mampu mempertahankan keunggulan . Jakarta , CNN Indonesia - - Timnas Argentina kembali gagal meraih kemenangan tanpa kehadiran Lionel Messi . Argentina untuk kali kedua beruntun meraih hasil imbang . Usai menerima umpan Miguel Trauco , Guerrero berhasil menahan bola dengan dada dan lolos dari kawalan Funes Mori.

B. Evaluation with ROUGE

Automatic evaluation is performed using the ROUGE-1, ROUGE-2, and ROUGE-L metrics to measure the similarity between the system-generated summaries and the reference summaries. The evaluation scores at each compression level are presented in Table 3.

TABLE III. EVALUATION RESULT WITH ROUGE

Compression Level	ROUGE-1	ROUGE-2	ROUGE-L
10%	0.3796	0.2459	0.3591
20%	0.4558	0.3154	0.4385
30%	0.4808	0.3433	0.4675

The highest ROUGE scores are recorded at the 30% compression level across all three metrics. A visualization of

the ROUGE evaluation results is presented in Figure 2 to provide a clearer illustration of the system's performance trends under each compression scenario.

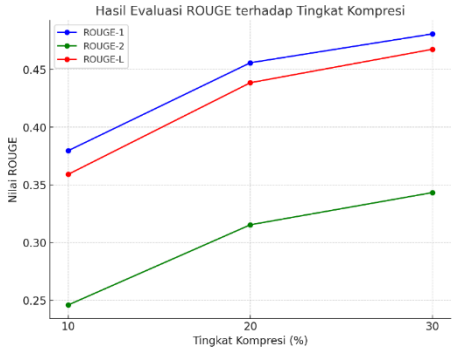

Fig. 2. ROUGE Evaluation Chart

C. Manual Evaluation

The results of the manual evaluation are visualized in bar charts shown in Figure. 3, Figure. 4, and Figure. 5, which represent the 10%, 20%, and 30% compression scenarios, respectively. Each chart illustrates the distribution of validator assessments across four aspects: content, relevance, redundancy, and specificity [20]. The evaluations were conducted by a validator with an academic background as a lecturer in Indonesian language, to ensure that the summaries were judged not only on informativeness but also on linguistic quality and readability.

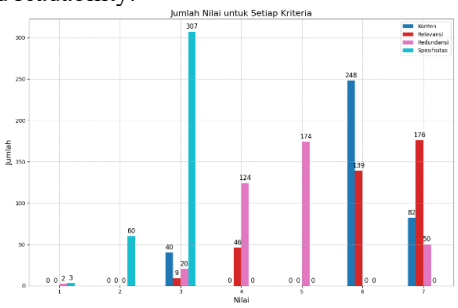

Fig. 3. Validator Assessment Results at 10%

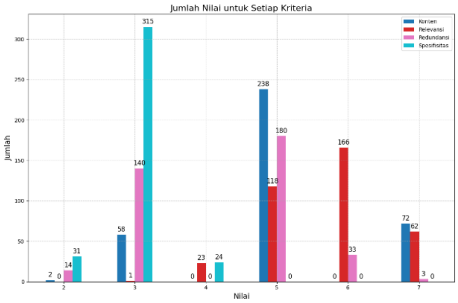

Fig. 4. Validator Assessment Results at 20%

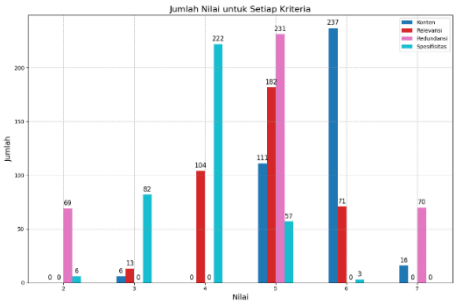

Fig. 5. Validator Assessment Results at 30%

D. Discussion

Based on the results of the automatic evaluation, increasing the summary compression level from 10% to 30% demonstrates an upward trend across all three ROUGE metrics. At 10% compression, the ROUGE-1, ROUGE-2, and ROUGE-L scores are 0.3796, 0.2459, and 0.3591, respectively. A significant improvement is observed at the 20% compression level, with ROUGE-1 reaching 0.4558, ROUGE-2 at 0.3154, and ROUGE-L at 0.4385. The highest scores are achieved at 30% compression, with ROUGE-1 at 0.4808, ROUGE-2 at 0.3433, and ROUGE-L at 0.4675. These results indicate that 30% compression yields the highest-quality summaries according to n-gram-based evaluation.

Manual evaluation conducted by a validator across four aspects content, relevance, redundancy, and specificity also supports these findings. At 10% compression, the content and relevance aspects receive relatively favorable ratings; however, scores for redundancy and specificity are low, indicating the presence of repetitive content and a lack of detail. At 20% compression, relevance improves while content slightly declines and specificity remains low. At 30% compression, all four aspects show consistent improvement, indicating that the summaries are more concise, relevant, and informative.

Overall, both automatic and manual evaluations confirm that the 30% compression level delivers the most optimal performance in producing high-quality summaries, both in terms of n-gram similarity and subjective content assessment.

IV. CONCLUSION

This study successfully applied a combination of the TextRank algorithm, Word2Vec, and Latent Semantic Analysis (LSA) for extractive summarization of Indonesian news articles. Automatic evaluation using ROUGE metrics showed that a 30% compression ratio yielded the best performance, with ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4808, 0.3433, and 0.4675, respectively. Manual evaluation by a validator also indicated that summaries at this compression level were more informative and relevant than those from other scenarios. These results demonstrate that the proposed approach effectively improves the quality of news text summarization.

REFERENCES

[1] N. D. Sukmono, "Clickbait Judul Berita Online dalam Pemberitaan Covid-19," Transformatika: Jurnal Bahasa, Sastra, dan Pengajarannya, vol. 5, pp. 1–13, Mar. 2021, doi: 10.31002/transformatika.v%vi%i.3643.

[2] N. Newman, R. Fletcher, K. Eddy, C. T. Robertson, and R. Kleis Nielsen, "Reuters Institute Digital News Report 2023," 2023, doi: 10.60625/risj-p6es-hb13.

[3] P. Majerczak and A. Strzelecki, "Trust, Media Credibility, Social Ties, and the Intention to Share Information Verification in an Age of Fake News," Behavioral Sciences, vol. 12, no. 2, Feb. 2022, doi: 10.3390/bs12020051.

[4] S. Thange, J. Dange, V. Karjule, and J. Sase, "A Survey on Text Summarization Techniques," International Journal of Scientific and Research Publications, vol. 13, no. 11, pp. 528–535, Nov. 2023, doi: 10.29322/ijsrp.13.11.2023.p14355.

[5] M. Xu, H. A. Rahman, and F. Li, "Text Summarization: A Bibliometric Study and Systematic Literature Review," Ingenierie des Systemes d'Information, vol. 29, no. 5, pp. 2077–2089, Oct. 2024, doi: 10.18280/isi.290538.

[6] J. Wijaya and A. S. Girsang, "Indonesian News Extractive Summarization using Lexrank and YAKE Algorithm," Statistics, Optimization and Information Computing, vol. 12, no. 6, pp. 1973–1983, Nov. 2024, doi: 10.19139/soic-2310-5070-1976.

[7] A. Rahmadhany, A. Aldila Safitri, and I. Irwansyah, "Fenomena Penyebaran Hoax dan Hate Speech pada Media Sosial," Jurnal Teknologi Dan Sistem Informasi Bisnis, vol. 3, no. 1, pp. 30–43, Jan. 2021, doi: 10.47233/jteksis.v3i1.182.

[8] K. Kurniawan and S. Louvan, "IndoSum: A New Benchmark Dataset for Indonesian Text Summarization," in Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 215–220. doi: 10.1109/IALP.2018.8629109.

[9] N. Basri and E. Utami, "Application of Word2Vec and LSTM Models in Sentiment Analysis of Mobile Legends User Reviews," SISTEMASI, vol. 14, p. 856, May 2025, doi: 10.32520/stmsi.v14i2.5074.

[10] P. L. Rodriguez and A. Spirling, "Word Embeddings What works, what doesn't, and how to tell the difference for applied research."

[11] D. F. Surianto, R. A. P. Kadir, F. Syafaat, M. M. Fakhri, and D. M. Rifqie, "Implementasi Metode Latent Semantic Analysis Pada Peringkasan Artikel Bahasa Indonesia Menggunakan Pendekatan Steinberger Jezek," JURIKOM (Jurnal Riset Komputer), vol. 9, no. 4, p. 894, Aug. 2022, doi: 10.30865/jurikom.v9i4.4620.

[12] D. Kurniadi, S. Farisa, C. Haviana, and A. Novianto, "Implementasi Algoritma Cosine Similarity pada sistem arsip dokumen di Universitas Islam Sultan Agung," TRANSFORMTIKA, vol. 17, no. 2, pp. 124–132, 2020, doi: 10.26623/transformatika.v17i2.1613.

[13] F. A. Nugroho, F. Septian, D. A. Pungkastyo, and J. Riyanto, "Penerapan Algoritma Cosine Similarity untuk Deteksi Kesamaan Konten pada Sistem Informasi Penelitian dan Pengabdian Kepada Masyarakat," Jurnal Informatika Universitas Pamulang, vol. 5, no. 4, p. 529, Dec. 2021, doi: 10.32493/informatika.v5i4.7126.

[14] K. Jane C. Patosa et al., "Enhancement of TextRank Algorithm using Coreference Resolution," International Journal of Research Publications, vol. 101, no. 1, May 2022, doi: 10.47119/ijrp1001011520223190.

[15] A. Kazemi, V. Pérez-Rosas, and R. Mihalcea, Biased TextRank: Unsupervised Graph-Based Content Extraction. 2020. doi: 10.18653/v1/2020.coling-main.144.

[16] M. Barbella and G. Tortora, "Rouge Metric Evaluation for Text Summarization Techniques," SSRN Electronic Journal, May 2022, doi: 10.2139/ssrn.4120317.

[17] Halimah, Surya Agustian, and Siti Ramadhani, "Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank," Jurnal CoSciTech (Computer Science and Information Technology), vol. 3, no. 3, pp. 371–381, Dec. 2022, doi: 10.37859/coscitech.v3i3.4300.

[18] A. Wahab, "Sampling dalam Penelitian Kesehatan," Jurnal Pendidikan dan Teknologi Kesehatan, vol. 4, pp. 38–45, May 2021, doi: 10.56467/jptk.v4i1.23.

[19] N. I. Majdina, B. Pratikno, and A. Tripena, "PENENTUAN UKURAN SAMPEL MENGGUNAKAN RUMUS," Jurnal Ilmiah Matematika dan Pendidikan Matematika (JMP), vol. 16, pp. 73–84, Jun. 2024, doi: 10.20884/1.jmp.2024.16.1.11230.

[20] J. Jia, L. Miratrix, B. Gawalt, B. Yu, and L. El Ghaoui, "Summarizing large-scale, multiple-document news data: sparse methods and human validation," Dec. 2013. [Online]. Available: http://statnews.org