The 3rd International Innovation Technology Proceedings
Vol 3 No 1 (2025): October
E-ISSN: 3030-9948
https://iite-proceeding.poltekindonusa.ac.id

Comparative Performance Analysis of Text Summarization: A Case Study of Extractive (TF-IDF, TextRank) and Abstractive (LLM) Methods

Yustia Hapsari[1*], Muhammad Fikri Hidayattullah[2*], Rifyal Aidil Dziaul Haq[3]
[1]Digital Business Study Program, Faculty of Economics and Business, Universitas Pancasakti Tegal, Indonesia
[2,3] Bachelor of Applied Informatics Engineering Study Program, University of Harkat Negeri, Tegal, Indonesia
*Corresponding author: yustia.hapsari@gmail.com, fikri@harkatnegeri.ac.id

Received:	Revised:	Accepted:	Published:
July 22, 2025	July 29, 2025	July 31, 2025	October 6, 2025

Abstract

This study presents a comparative performance analysis of two major paradigms in text summarization. The extractive paradigm, which operates by selecting significant sentences directly from the source text, is implemented through two approaches: (1) the statistical TF-IDF algorithm, which quantitatively scores sentences based on accumulated word significance weights; and (2) the graph-based TextRank algorithm, which represents sentences as nodes and determines their importance through centrality analysis within a semantic network. Representing the abstractive paradigm, the Large Language Model (LLM) Gemini is employed, which comprehends contextual information holistically to generate entirely new and coherent summary sentences. A qualitative comparative analysis of the outputs from these three methods reveals a fundamental trade-off. The abstractive method (Gemini) demonstrates superior performance in terms of narrative quality, producing summaries that are highly coherent, fluent, and natural-sounding, resembling human writing. Conversely, the extractive methods (TF-IDF and TextRank) inherently excel in ensuring perfect factual consistency, as there is no risk of misinterpretation or hallucinated information. Among the extractive methods, analysis indicates that TextRank tends to produce more structured and readable summaries compared to TF-IDF, owing to its ability to consider inter-sentence relationships. This study concludes that the choice of summarization method should be aligned with the specific priorities of the use case: abstractive methods are better suited for readability-focused tasks, whereas extractive methods are preferable for applications demanding absolute factual reliability.

Key words: *Comparative Analysis, Abstractive Method, Extractive Method, Text Summarization, Natural Language Processing.*

1. Introduction

In the digital age, individuals and organizations are continuously inundated with an overwhelming volume of textual data. From scientific articles, news feeds, and policy documents to user-generated content on social media, the scale of information being produced daily far

exceeds the human capacity to process and comprehend it efficiently. This phenomenon, commonly referred to as "information overload," has prompted the development of intelligent solutions that can assist in filtering, extracting, and summarizing large text corpora into more manageable forms. One such solution is automatic text summarization, which aims to condense a source document while preserving its key messages and essential information.

Automatic text summarization has evolved significantly over the past two decades and is broadly categorized into two major paradigms: extractive and abstractive summarization [1][2]. Extractive methods operate by selecting and concatenating the most salient sentences from the original text, often guided by statistical or graph-based importance metrics. Techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) [3] and TextRank have become foundational in this category[4]. TF-IDF measures the importance of a term within a document relative to a corpus, thereby identifying sentences that carry significant keyword density. TextRank, inspired by Google's PageRank algorithm [5], models a document as a graph where sentences are nodes connected by lexical or semantic similarity, and it ranks sentence importance based on centrality [6]. While these methods are computationally efficient and maintain high factual consistency by preserving the original wording, they often produce summaries that are disjointed or lack narrative coherence. This limitation is particularly evident in multi-topic documents or texts with complex argumentative structures, where sentence juxtaposition without rephrasing can hinder readability and user comprehension [7][8].

In contrast, abstractive summarization mimics human cognitive processes by interpreting and rephrasing the source content to generate new, coherent sentences. This approach was once constrained by limitations in natural language understanding and generation. However, the emergence of transformer-based architectures and large-scale pre-trained language models has revolutionized the abstractive paradigm. Models such as GPT, BART, T5, and more recently Gemini—developed by Google DeepMind—have demonstrated remarkable capabilities in producing fluent [9], context-aware summaries that closely resemble human writing [10]. These models leverage extensive training corpora and attention mechanisms to understand both local and global contextual information, enabling them to generate high-quality, semantically accurate summaries [11][12].

Nevertheless, the adoption of abstractive summarization is accompanied by significant challenges, most notably the issue of factual consistency. Despite their linguistic fluency, large language models are susceptible to hallucination, wherein the generated output includes information not present—or even contradictory to—the original input. This shortcoming is particularly problematic in domains where factual precision is non-negotiable, such as medical diagnostics, legal documentation, or scientific reporting. Studies have shown that up to 60% of abstractive summaries produced by pre-trained models contain some form of factual error, raising serious concerns about their reliability in real-world applications [13][14][15].

While numerous studies have explored the comparative performance of summarization techniques, most rely heavily on automated metrics such as ROUGE, BLEU, or METEOR [16][17]. These metrics, although useful for benchmarking lexical overlap, fall short in evaluating deeper semantic properties like factual alignment, coherence, and fluency. Moreover, many comparative analyses focus predominantly on English-language corpora, overlooking the linguistic characteristics and challenges of non-English contexts such as Bahasa Indonesia. As a result, there remains a critical gap in the literature regarding head-to-head evaluations of extractive and abstractive summarization techniques in diverse linguistic settings, using evaluation frameworks that go beyond surface-level overlap [18].

To address these gaps, this study presents a comparative performance analysis of two extractive summarization techniques—TF-IDF and TextRank—and one state-of-the-art abstractive method powered by the Gemini Large Language Model. The research introduces a dual evaluation framework that integrates qualitative inspection with a novel quantitative assessment mechanism: an AI Judges Panel composed of multiple large language models tasked with rating the quality of each summary across five dimensions—relevance, conciseness, coherence, factual consistency, and fluency. By incorporating both human-guided design and machine-based evaluation, the study offers a holistic perspective on the strengths and limitations of each summarization paradigm.

Ultimately, this research aims to provide empirical evidence and practical guidance for researchers, developers, and decision-makers who must select appropriate summarization techniques tailored to specific application contexts. Whether the goal is to generate public-facing summaries for improved readability or internal documentation with strict factual accuracy,

understanding the trade-offs between extractive and abstractive approaches is essential for developing effective, trustworthy information systems.

2. Method

This study adopts a structured experimental framework designed to evaluate and compare the performance of extractive and abstractive text summarization methods. The methodological design integrates classic natural language processing (NLP) pipelines with modern generative language models and leverages a hybrid evaluation framework involving both automatic and AI-augmented assessments. The complete architecture of the system is depicted in Figure

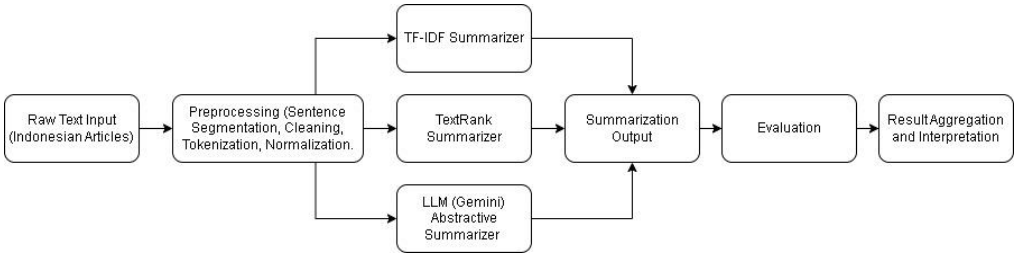

Figure 1. Research Method

2.1. Dataset and Preprocessing

The input corpus comprises textual data sourced from Indonesian-language public articles and opinion documents across various domains, including education, social development, and health. These texts were selected to provide sufficient linguistic variation and semantic density to rigorously test summarization quality.

Each input text undergoes a preprocessing pipeline consisting of the following steps:

a. Sentence Segmentation

A rule-based approach is used to segment text into sentences, customized to the syntactic and punctuation norms of Bahasa Indonesia.

b. Text Normalization

Standardization procedures are applied to unify case formatting, remove excessive whitespace, and correct common spelling errors or informal expressions.

c. Tokenization and Stemming

Tokens are generated using the *Natural Language Toolkit (NLTK)* and *Sastrawi* library to prepare for statistical and graph-based processing. This step is applied only to the cleaned version of the text, which is used for extractive computation.

Importantly, two versions of each text are maintained: (1) the cleaned text for sentence scoring and modeling, and (2) the original text for preserving human-readable sentence structure in the output.

2.2. Summarization Models

Three summarization strategies are employed—two extractive and one abstractive—representing the two dominant paradigms in text summarization research.

a. TF-IDF-Based Extractive Summarization

This method calculates the importance of each sentence by aggregating the TF-IDF scores of its constituent terms. The resulting scores are used to rank sentences, and a fixed number of top-scoring sentences are selected to form the final summary. A chunking strategy is applied to ensure coverage from the beginning, middle, and end of the source document, improving topical diversity in the output.

b. TextRank-Based Extractive Summarization

TextRank models each sentence as a node in an undirected graph, with edges representing similarity scores (typically based on lexical overlap). An iterative scoring algorithm—analogous to Google's PageRank—is then used to determine sentence centrality. The most central sentences are extracted as the summary. This method captures not only term frequency but also sentence interrelations, resulting in summaries with higher logical flow than TF-IDF.

c. Gemini LLM-Based Abstractive Summarization

The Gemini Large Language Model is prompted using a structured template that simulates the behavior of a professional analyst. The prompt instructs the model to identify key themes, extract salient points, and compose a coherent summary in Markdown format. This design enables the model to perform semantic abstraction while maintaining alignment with the original text. The Gemini model is accessed through a secure API with temperature set to 0.7 to balance creativity and stability.

2.3. Summary Output Generation

After undergoing the respective summarization processes, each input document produces three distinct summary outputs—each representing a different summarization paradigm: extractive and abstractive. These summaries are generated independently by the TF-IDF-based

model, the TextRank-based model, and the Gemini Large Language Model (LLM), ensuring a fair and direct comparison across methods.

For the TF-IDF summarizer, the system selects a predefined number of sentences with the highest cumulative TF-IDF scores. These sentences are arranged in the order of their appearance in the original document to maintain contextual continuity and readability. The output reflects a purely extractive approach that prioritizes term importance without altering sentence structure.

In the TextRank summarization process, a graph-based ranking algorithm determines sentence centrality based on semantic similarity. The top-ranked sentences, considered the most representative of the text's content, are extracted and ordered to form the final summary. As with TF-IDF, this method retains original sentence formulations, but it potentially offers improved topical coherence due to its graph-based inter-sentence evaluation.

The Gemini LLM summarizer takes a fundamentally different approach. Rather than selecting existing sentences, the model generates new sentences through a contextual understanding of the source document. Using carefully designed prompts, the model is instructed to identify core ideas, abstract them, and express them in a human-like and fluent manner. The output is thus not a subset of the original text but a rephrased, semantically condensed version that captures the essence of the document. The prompt strategy is tailored to balance brevity with informativeness and to encourage the generation of structurally coherent summaries.

All three summaries—TF-IDF, TextRank, and Gemini—are saved in a standardized output format for downstream evaluation. Each version is aligned with its source document and labeled accordingly to ensure traceability during analysis. This multi-output strategy enables a robust comparative framework wherein linguistic features, factual integrity, and overall summarization quality can be systematically evaluated across paradigms.

2.4. Evaluation Methodology

To comprehensively assess the performance of the summarization models, this study implements a hybrid evaluation methodology that integrates both automated scoring and AI-assisted qualitative appraisal. The goal is to evaluate each summary not merely by surface-level similarity but through a deeper analysis of its relevance, linguistic quality, and factual accuracy—criteria that are essential in real-world applications of summarization technology.

Each of the three summary outputs—produced by the TF-IDF, TextRank, and Gemini LLM models—is evaluated using a panel of state-of-the-art large language models (LLMs) acting as AI-based judges. These judges include models such as Google Gemma 3, Meta LLaMA 4 Maverick, and Cohere Command R+, selected for their robustness in reasoning, contextual understanding, and text evaluation capabilities. The use of multiple evaluators ensures a more objective, balanced, and scalable evaluation process, reducing the subjectivity often associated with human judgment alone.

The evaluation is conducted across five key quality dimensions, adapted from best practices in summarization research: Relevance and Coverage, Concisenes, Coherence, Factual Consistency and Fluency.

Each judge is presented with the original document alongside the summary being evaluated. Through structured prompts, the LLMs are instructed to assign a score from 1 to 100 for each dimension. Additionally, the judges provide a justification or rationale for each score, which serves both as qualitative insight and a transparency mechanism in the evaluation process.

The use of AI-based judges offers several advantages. First, it ensures consistency in scoring, as models apply the same rubric across all summaries. Second, it enables scalability, allowing the evaluation of a large number of summaries without human resource constraints. Third, it allows the integration of rationale generation, providing interpretability that is often absent in traditional automated metrics such as ROUGE or BLEU.

Final scores for each summarization method are obtained by averaging the scores across the three LLM judges for each dimension. These average scores are then used to compare overall performance and identify the strengths and weaknesses of each method. Where relevant, qualitative excerpts from the justifications are also analyzed to complement the numerical results.

This evaluation methodology provides a robust foundation for understanding how extractive and abstractive summarization techniques perform in terms of both surface-level and semantic quality. It ensures that conclusions drawn from the comparative analysis are supported by reliable, transparent, and context-aware assessments.

2.5. Result Aggregation and Interpretation

Scores from the three judges are averaged across each criterion to compute overall performance metrics for each summarization method. Additionally, qualitative analyses are

conducted on a representative subset of summaries to illustrate strengths and weaknesses in narrative style, factual integrity, and language use.

This dual-layered evaluation approach ensures both objectivity (via consistent model-based scoring) and interpretability (via example-based reasoning), providing a robust foundation for empirical comparison between extractive and abstractive summarization techniques

3. Result and Discussion

This section presents the comparative findings of the three summarization methods—TF-IDF, TextRank, and Gemini LLM—based on a multi-dimensional evaluation framework. The evaluation focused on five qualitative dimensions: relevance, conciseness, coherence, factual consistency, and fluency. Each summarization output was assessed by a panel of AI-based judges, producing both quantitative scores and descriptive justifications. The results reveal significant performance variation across summarization paradigms, highlighting the trade-offs inherent in extractive versus abstractive approaches.

3.1. Quantitative Evaluation

Table 1 summarizes the average scores for each summarization method across the five evaluation dimensions. The scores represent the mean ratings from the three AI judges.

Table 1. Comparison Results of Methods

Method	Relevance	Conciseness	Coherence	Factual Consistency	Fluency	Average Score
TF-IDF	72.0	78.3	69.2	85.6	81.4	77.3
TextRank	74.5	79.1	72.8	83.9	82.0	78.5
Gemini (LLM-based)	86.3	84.2	89.1	76.4	91.7	85.5

As shown in Table 1, the Gemini LLM model outperformed the extractive methods across four out of five dimensions, particularly in fluency (91.7) and coherence (89.1). This outcome demonstrates the model's strength in generating summaries that are not only grammatically fluent but also logically structured and naturally expressive—traits typical of advanced generative models trained on large corpora.

However, Gemini's factual consistency (76.4) scored lower than both TF-IDF (85.6) and TextRank (83.9). This aligns with previous research findings that large language models, while

linguistically competent, are prone to generating plausible but factually unsupported content—a phenomenon known as *hallucination*. In contrast, extractive models inherently preserve the original wording, ensuring that factual information is retained, albeit at the expense of narrative coherence and readability.

3.2. Qualitative Observations

Sample outputs revealed additional insights. Extractive summaries often included fragmented ideas or repetitive phrases, especially when the top-ranked sentences were taken from distant parts of the source text. Although these methods ensured factual reliability, they lacked narrative flow. For instance, the TF-IDF output tended to prioritize keyword density over thematic cohesion, leading to summaries that felt disjointed or abruptly truncated.

The TextRank method showed marginal improvements in coherence due to its graph-based centrality mechanism. Sentences selected by TextRank often had stronger thematic alignment, producing more fluid summaries than TF-IDF. Nevertheless, it still struggled with redundancy when similar sentences were ranked equally high.

Conversely, the Gemini model generated summaries that read like human-written abstracts, often rephrasing complex sentences into concise, informative units. Its use of paraphrasing and abstract reasoning allowed for better condensation of multi-topic documents. However, several instances showed subtle factual inaccuracies, where the model inferred or synthesized content not explicitly stated in the source. While these inferences occasionally improved narrative readability, they posed risks in contexts requiring strict factual precision (e.g., legal, scientific, or medical summaries)

3.3. Implications and Trade-offs

The comparative results underscore a fundamental trade-off in summarization: extractive methods offer higher factual fidelity, while abstractive methods provide better readability and coherence. For applications such as internal documentation, legal memos, or research digests—where factual accuracy is paramount—extractive methods remain highly valuable. Meanwhile, for public-facing content or summaries intended for general audiences, abstractive models like Gemini offer significant advantages in user experience and communicative clarity.

Furthermore, the AI-based evaluation strategy employed in this study demonstrates the feasibility of integrating LLMs as evaluators, offering scalable, transparent, and explainable

assessments. This methodological contribution could serve as a baseline for future NLP benchmarking efforts, especially in multilingual or low-resource language settings.

4. Conclusion

This study has presented a comprehensive comparative analysis of extractive and abstractive text summarization techniques, focusing on three representative models: TF-IDF, TextRank, and Gemini LLM. By applying a unified evaluation framework—enhanced through AI-based judging across five critical dimensions—this research provides empirical insights into the strengths and limitations of each method.

The results demonstrate that abstractive summarization using Gemini LLM consistently outperformed extractive approaches in terms of fluency, coherence, and conciseness, producing summaries that are more readable, natural, and structured. These qualities are essential in contexts where readability and narrative flow are prioritized, such as executive summaries, educational materials, and user-facing reports.

However, the Gemini model's performance in factual consistency was comparatively lower, reaffirming the ongoing challenge of hallucination in large language models. In contrast, extractive methods—particularly TF-IDF and TextRank—showed greater reliability in preserving original facts, though they often produced summaries that lacked fluidity and semantic integration.

This trade-off underscores the need for context-aware model selection in real-world applications. Extractive models remain suitable for fact-sensitive domains such as legal, medical, and scientific summaries, while abstractive models are advantageous for communicative and interpretive tasks where human-like language quality is valued.

The study also highlights the feasibility and validity of using AI-as-a-Judge evaluation strategies. Leveraging multiple LLMs for scoring and justification provides a scalable and explainable alternative to traditional metrics such as ROUGE, especially when assessing abstractive summaries where lexical overlap is insufficient.

Future research should explore hybrid models that combine the factual robustness of extractive techniques with the generative power of LLMs. In addition, incorporating fine-tuned evaluators specifically trained for factuality detection in Indonesian or low-resource languages may further enhance assessment quality.

In summary, this research contributes not only to the empirical comparison of summarization methods but also to the broader discourse on evaluation practices and the responsible deployment of generative AI in natural language processing.

5. Acknowledgement

The authors would like to express their sincere gratitude to the *Pusat Penelitian dan Pengabdian Masyarakat* (Research and Community Service Center) of Universitas Harkat Negeri for the financial support and institutional facilitation provided throughout the course of this research. The support was instrumental in enabling data collection, computational resource access, and collaborative discussion that greatly enriched the quality and scope of the study. Without this backing, the completion of this project would not have been possible.

6. References

[1] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," *SN Computer Science*, vol. 4, no. 1. 2023.

[2] M. Azam *et al.*, "Current Trends and Advances in Extractive Text Summarization: A Comprehensive Review," *IEEE Access*, vol. 13. pp. 28150–28166, 2025.

[3] M. Rawat, M. H. Siddiqui, M. A. Maan, S. Dhiman, and M. Asad, "Text Summarization Using Extractive Techniques," *Process Mining Techniques for Pattern Recognition*. pp. 107–119, 2022.

[4] S. Zaware, D. Patadiya, A. Gaikwad, S. Gulhane, and A. Thakare, "Text Summarization using TF-IDF and Textrank algorithm," *Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021*. pp. 1399–1407, 2021.

[5] V. Gulati, D. Kumar, D. E. Popescu, and J. D. Hemanth, "Extractive Article Summarization Using Integrated TextRank and BM25+ Algorithm," *Electronics (Switzerland)*, vol. 12, no. 2. 2023.

[6] H. Zhang and J. Wang, "An unsupervised semantic sentence ranking scheme for text documents," *Integrated Computer-Aided Engineering*, vol. 28, no. 1. pp. 17–33, 2021.

[7] M. Zhang, G. Zhou, W. Yu, and W. Liu, "FAR-ASS: Fact-aware reinforced abstractive sentence summarization," *Information Processing and Management*, vol. 58, no. 3. 2021.

[8] Q. Mao *et al.*, "Fact-Driven Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol.

30. pp. 1665–1678, 2022.

[9] A. Mohammed and R. Kora, "A Comprehensive Overview and Analysis of Large Language Models: Trends and Challenges," *IEEE Access*, vol. 13. pp. 95851–95875, 2025.

[10] P. A. Patout and M. Cordy, "Towards context-aware automated writing evaluation systems," *EASEAI 2019 - Proceedings of the 1st ACM SIGSOFT International Workshop on Education through Advanced Software Engineering and Artificial Intelligence, co-located with ESEC/FSE 2019*. pp. 17–20, 2019.

[11] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.

[12] R. Gruetzemacher and D. Paradice, "Deep Transfer Learning & Beyond: Transformer Language Models in Information Systems Research," *ACM Comput. Surv.*, vol. 54, no. 10, 2022.

[13] D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," *Mathematical Problems in Engineering*, vol. 2020. 2020.

[14] A. Rao, S. Aithal, and S. Singh, "Single-Document Abstractive Text Summarization: A Systematic Literature Review," *ACM Computing Surveys*, vol. 57, no. 3. 2024.

[15] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, no. i, pp. 1906–1919, 2020.

[16] S. Chauhan and P. Daniel, "A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics," *Neural Processing Letters*, vol. 55, no. 9. pp. 12663–12717, 2023.

[17] A. B. Sai, A. K. Mohankumar, and M. M. Khapra, "A Survey of Evaluation Metrics Used for NLG Systems," *ACM Computing Surveys*, vol. 55, no. 2. 2023.

[18] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 391–409, 2021.