EDUMATIC Jurnal Pendidikan Informatika

e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Volume 10, Issue 1, **30-39.**
https://e-journal.hamzanwadi.ac.id/index.php/edumatic
DOI: 10.29408/edumatic.v10i1.33443

Feature Interaction and Performance Analysis of RankSum-Based Extractive Summarization in Indonesian Scientific Articles

Verrino Adityya [1,*] , Yohannes [1]

[1] Universitas Multi Data Palembang, Indonesia
[*] Correspondence: Verrino Adityya, Universitas Multi Data Palembang, Indonesia
✉ verrinoadityya_2226250053@mhs.mdp.ac.id

Copyright: © 2026 by the authors

Received: 11 December 2025 | Revised: 3 January 2026 | Accepted: 25 January 2026 | Published: 3 March 2026

Abstract
The extractive summarization of Indonesian scientific articles is hindered by a domain mismatch where established methodologies rely on news-corpus assumptions, whereas Indonesian scientific discourse follows rigid, IMRaD-driven structural and lexical patterns. This study aims to systematically analyze feature interaction effects and saturation behaviour in RankSum-based extractive summaries for Indonesian scientific articles. Designed as a controlled comparative experiment, this research evaluates a RankSum framework integrating variables, such as graph-based, semantic-thematic vectors, and structural heuristics. The dataset comprises 2,897 Indonesian journal articles (2021-2025) collected via web scraping from open-access university repositories. Analysis across 31 scenarios demonstrates that for Indonesian scientific articles, the assumption that increasing feature density improves performance is flawed, instead a feature saturation effect occurs. Results show that a 4-feature combination maximizes unigram lexical precision (ROUGE-1 0.3564), whereas the full 5-feature fusion is necessary to preserve global semantic integrity, structural flow, and stable (ROUGE-L 0.2018; BERTScore 0.6977). This study establishes a generalizable principle for domain-aware ATS by demonstrating that overcoming domain mismatch relies on navigating feature saturation through selection aligned with the document's inherent logic rather than raw feature quantity.

Keywords: extractive text summarization; indonesian scientific articles; multi-feature integration; rank fusion; rouge-bertscore evaluation

To cite this article: Adityya, V., & Yohannes, Y. (2026). Feature Interaction and Performance Analysis of RankSum-Based Extractive Summarization in Indonesian Scientific Articles. *Edumatic: Jurnal Pendidikan Informatika*, *10*(1), 30–39. https://doi.org/10.29408/edumatic.v10i1.33443

INTRODUCTION

The global surge in scientific articles has shifted the challenge from information accessibility to information filtration. In Indonesia, the proliferation of open-access journals has created a massive corpus that researchers can no longer navigate manually. The manual reading process, which is time-consuming, has the potential to cause the loss of critical information (Bandaru & Radhika, 2022). While automatic text summarization (ATS) offers a computational solution, the field is currently constrained by a significant domain mismatch. Most established ATS methodologies are rooted in news-corpus assumptions that falter when applied to the dense, terminological, and hierarchical structure of scientific articles. News context allows for simple lead-based extraction. In scientific journals, information is distributed

Edumatic: Jurnal Pendidikan Informatika is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

across rigid discursive deductive-inductive patterns, results-heavy conclusions, and specialized keyword densities, which require a more nuanced, domain-aware approach to feature selection.

Recent advancements in ATS distinguish between abstractive and extractive methods. Abstractive models, while capable of human-like synthesis, are notoriously susceptible to hallucinations, often generating information that deviates from the source text (Zhou et al., 2025). In scientific domains, where factual fidelity is non-negotiable, extractive methods remain the gold standard due to their ability to preserve original phrasing. However, popular single-feature algorithms like TextRank rely heavily on graph-based lexical centrality, often ignoring the semantic depth and structural heuristics inherent in scientific writing (Cai et al., 2022).

The emergence of rank fusion approaches, RankSum (Joshi et al., 2022), suggested that aggregating multiple feature dimensions could mitigate the limitations of single-feature ATS. This study leverages rank fusion to focus on domain adaptation for Indonesian scientific articles, the interaction analysis of diverse feature sets, and the empirical identification of feature saturation. However, current literature on Indonesian text summarization remains dominated by single-approach methodologies that lack the multidimensionality required for complex scientific texts. Previous studies have primarily utilized graph-based centrality like TextRank (Gulati et al., 2023; Wijaya & Girsang, 2024), topic modelling and MMR (Faisal et al., 2024), or Ant System (Girsang & Amadeus, 2023). Furthermore, while supervised method LSTM (Fitrianah & Jauhari, 2022) have been explored, their dependency on massive labeled datasets remains a bottleneck for many Indonesian scientific sub-domains. Even in global contexts, multi-feature approaches remain relatively limited to specific extractive (Aziz et al., 2025; Joshi et al, 2023; Liu et al., 2024; Yulianti et al., 2023), abstractive frameworks (Aurelia et al., 2024; Onan & Alhumyani., 2024; Ma et al., 2022; Ulker & Ozer, 2024), and hybrid (Lau & Tan, 2024; Zhang et al., 2024), with most of the research still anchored to English news corpora like CNN/DailyMail or DUC. To date, no study has systematically examined the interaction effects and trade-offs of multidimensional rank fusion features in Indonesian scientific articles.

Current research is not merely a lack of application, but a lack of understanding regarding how ranking features behave and interact within this specific Indonesian linguistic and structural ecosystem. Without a domain-specific analysis, the development of ATS systems risks adopting suboptimal or misleading architectures that prioritize feature quantity over feature alignment. There is currently no empirical evidence exploring the trade-offs between lexical precision and semantic coherence when rank fusion is applied to Indonesian scientific texts.

This study addresses by evaluating a RankSum framework optimized for the Indonesian scientific domain. We integrate five distinct features, TextRank, Semantic Rank, Topic Rank, Keyword Count Rank, and Position Rank. Instead of prioritizing raw system metrics, this study aims to provide a conceptual understanding of the trade-offs between lexical precision and semantic coherence when rank fusion is applied to the IMRaD-driven (Introduction, Methods, Results, and Discussion) structure of Indonesian scientific articles. The primary contributions include empirical evidence that lean feature combinations can outperform full feature fusion in scientific contexts for feature saturation effect, identification features interaction of the critical trade-offs between lexical and semantic metrics, and a demonstration of RankSum's domain-sensitivity, providing a new empirical framework for domain-aware NLP applications.

METHOD

This study is designed as a controlled comparative experiment. This aims to test specific feature interactions and the behavioral effectiveness of various rank fusion scenarios in the Indonesian scientific article domain. To achieve this, we evaluate how different combinations of graph-based, semantic, and structural features influence summary quality. The data used for this study is a dataset of Indonesian scientific journals in PDF format, collected through Selenium on open-access journal repositories in the Palembang region.

While the repositories were localized, Indonesian scientific discourse follows standardized structural and lexical patterns. To ensure generalization, the corpus included across 12 diverse disciplines. The corpus of 2.897 journal documents was partitioned into an 80% development set and a 20% testing set. To ensure robust performance across different data subsets and optimize ranking weights, k-fold cross-validation was implemented on the development set, utilizing a 60/20 training and validation within each fold. Preprocessing converted PDF files into structured text. The abstract was extracted as the reference summary. The article content and Keywords were extracted as primary inputs.

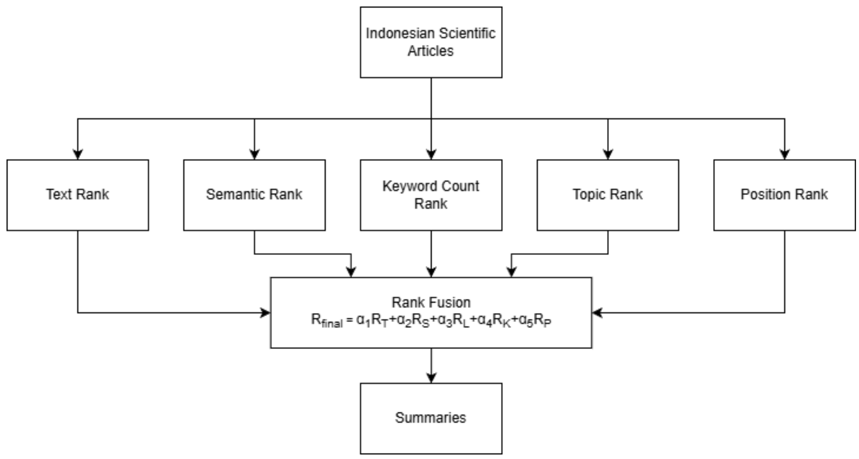

Figure 1. RankSum architecture

The RankSum architecture in Figure 1 is grounded in a multidimensional salience framework. We integrate five features across three salience dimensions. Graph salience via TextRank to capture lexical connectivity by modeling sentence centrality. Semantic-topic salience via Semantic employs IndoSBERT model to measure sentences to vectors and cosine similarity for a document centroid, which serves as a global representation, and Topic utilizes Latent Dirichlet Allocation (LDA) to provide a thematic signal and as a supporting signal. Potential noise is anticipated and controlled through the rank fusion process and the data-driven weighting derived from empirical training in LDA. Structural-lexical salience via Keyword Count Rank to track the frequency of keywords. Sentences with high keyword density are reliable indicators of subject-matter relevance (Liu, 2024) in the Indonesian scientific domain. Position Rank utilizes heuristic weights that reflect the IMRaD structure common to scientific journals. Critical insights and research findings are concentrated in initial sentences of a paragraph or the summaries found in the results or conclusion.

$$a_i = \frac{1}{3}\left(\frac{avg_{ROUGE-1}}{sum_{ROUGE-1}} + \frac{avg_{ROUGE-2}}{sum_{ROUGE-2}} + \frac{avg_{ROUGE-L}}{avg_{ROUGE-L}}\right) \tag{1}$$

$$R_{final} = a_1R_T + a_2R_S + a_3R_L + a_4R_K + a_5R_P \tag{2}$$

Scores are normalized using Min-Max Normalization. The weight for each ranking feature is calculated using equation (1), where a_i is the weight for each feature. The final RankSum score is calculated using equation (2). To ensure comparison across 31 scenarios, a compression rate of 20% was applied to generate all summaries (Gulati et al., 2023). While this ratio is heuristic and known to impact performance metrics significantly, optimizing the compression ratio itself is beyond the current scope, which focuses on feature saturation. Redundancy was mitigated through a Sentence Novelty Extractor using bigram and trigram matching. Evaluation was conducted using ROUGE to assess lexical which measures the n-gram overlap or the longest common sequence of the summaries via ROUGE-1, ROUGE-2, and ROUGE-L scores. BERTScore tests the model's ability to generalize semantically beyond surface-level word matching (Azam et al., 2025). By decoupling the optimization target (lexical) from the generalization test (semantic), we mitigate the risk of metric-specific bias while identifying the functional trade-offs of the fusion.

RESULTS AND DISCUSSION

Results

Experiments were conducted on the test data to evaluate the effectiveness of the rank fusion approach relative to baseline methods and multi-feature variations. The primary results involve ROUGE-1, ROUGE-2, and ROUGE-L F1 scores for lexical overlap, alongside BERTScore for semantic relevance. This evaluation phase utilized 31 distinct combination scenarios, ranging from single-feature baselines to a complete 5-feature fusion. Crucially, these numerous scenarios are not merely technical explorations but serve to reveal systematic patterns of feature behavior and their interactions within rank fusion. By analyzing this, the study aims to discover general empirical principles of Indonesian scientific summarization rather than merely identifying a singular best model. The weight stability analysis in Figure 2 reveals a nearly uniform distribution of feature weights across the tested scenarios. These weights are relatively stable. This suggests that the variation in performance across scenarios stems primarily from feature presence (categorical suitability) rather than weight dominance, confirming that each salience dimension provides a unique, non-redundant contribution to the summary.

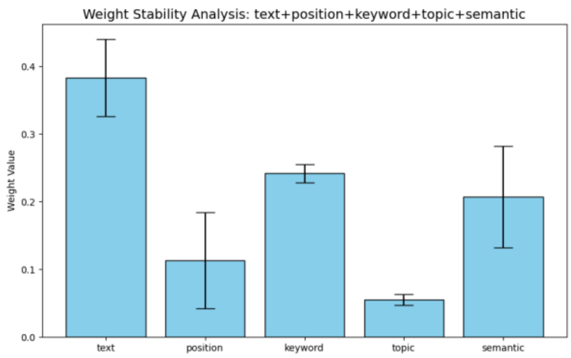

Figure 2. Weight stability showing average feature weights and standard deviation across cross-validation folds

The weights for each feature used in every combination were determined using equation (1), and each weight was applied in equation (2). The resulting performance metrics for each combination are detailed in Table 1. These test evaluation results cover ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore, providing a comprehensive view of how different feature sets perform relative to a standard Lead-3 baseline. This baseline was incorporated to provide a standard for extractive performance. The results demonstrate that all RankSum scenarios significantly and consistently outperform the Lead-3 baseline across every tested metric. This substantial performance gain confirms that the rank fusion framework is not merely numerically superior but provides a statistically robust improvement for the Indonesian scientific article domain. The consistent gap between the RankSum and the baseline highlights that domain-aware feature selection is a fundamental requirement for scientific summarization.

Table 1. Test evaluation results

Combination	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
text, position, keyword, topic, semantic	0.356461	0.134215	0.201834	0.697704
text, keyword, semantic	0.355253	0.137161	0.201739	0.697232
position, keyword, topic, semantic	0.354029	0.132755	0.200496	0.697463
text, keyword	0.350062	0.137646	0.200494	0.695913
text. keyword. topic	0.348717	0.134925	0.200155	0.696055
text. keyword. topic. semantic	0.356827	0.134380	0.199730	0.697668
text. position. keyword. semantic	0.352552	0.132105	0.198913	0.696575
position. keyword. semantic	0.352437	0.131520	0.198661	0.697160
text. position. keyword	0.339039	0.130479	0.197825	0.695644
keyword. semantic	0.353724	0.131448	0.197630	0.697330
keyword. topic. semantic	0.353177	0.130147	0.197449	0.696702
text. position. keyword. topic	0.340017	0.128747	0.196267	0.695487
position. keyword. topic	0.334523	0.126414	0.192426	0.694354
position. keyword	0.327687	0.124393	0.191127	0.692905
text. semantic	0.341278	0.123959	0.190849	0.692707
text. topic. semantic	0.345280	0.121941	0.190137	0.693156
text. position. semantic	0.332166	0.118631	0.189251	0.691816
keyword. topic	0.328670	0.121325	0.188638	0.691135
text. position. topic. semantic	0.332476	0.116888	0.186984	0.693186
topic. semantic	0.341603	0.115917	0.186131	0.692562
position. topic. semantic	0.337014	0.114027	0.186040	0.693135
text. position. topic	0.315950	0.116239	0.184816	0.687832
position. semantic	0.330301	0.110531	0.183988	0.691130
text. position	0.306389	0.111589	0.179886	0.684958
text. topic	0.307151	0.110715	0.177030	0.682079
position. topic	0.264565	0.078315	0.152473	0.676637
Baseline (Lead-3)	0.220249	0.058194	0.131881	0.662385

The evaluation reveals that the combination of TextRank. Keyword Count. Topic Rank. and Semantic Rank attained the highest ROUGE-1 score (0.3568). slightly outperforming the model incorporating all five features (0.3564). This performance highlights the structured nature of scientific articles. where keyword repetition and sentence position act as primary salience indicators. reinforcing the necessity of domain-specific feature selection. Regarding ROUGE-2 performance. which measures phrase-level bigram overlap. the pairing of TextRank and Keyword Count achieved the highest score of 0.1376. However. the consistent improvement observed when incorporating Semantic Rank suggests a shift from surface-level lexical salience toward deeper semantic coherence. For instance. adding semantic features to the text. position. and keyword configuration increased the ROUGE-2 score from 0.1304 to 0.1321. demonstrating that semantic signals capture structural logic and phrase-level meaning often overlooked by purely statistical or graph-based methods.

The comprehensive five-feature fusion (text. position. keyword. topic. semantic) remains the most effective configuration and the most stable one. achieving a peak performance of 0.2018 and 0.6977 at ROUGE-L and BERTScore. while still achieving average ROUGE-1 0.3565 and ROUGE-2 0.1342. Notably. the transition from the best-performing four-feature (position. keyword. topic. semantic) to the full five-feature model yielded a marginal gain 0.0013 in ROUGE-L and 0.0002 in BERTScore. This evidence of a performance confirms a state of feature saturation. where the incremental contribution of the fifth feature (text) is statistically minimal despite achieving the absolute peak.

The elevated ROUGE-L score reflects the structural suitability of the generated summaries compared to human-written abstracts. which adhere to a rigid logical sequence within the IMRaD framework. This performance is further validated by the high BERTScore. which indicates that the model captures deep contextual and semantic similarity rather than just surface-level word overlap. While ROUGE-L confirms that the system successfully replicates the sequential information flow and structural alignment of scientific inquiry. the BERTScore suggests that the selected sentences maintain the original meaning and nuanced technical concepts through dense vector representations. By prioritizing sentence position and lexical centrality. the system ensures that the summaries are both structurally coherent and semantically equivalent to the source text.

Further qualitative failure analysis reveals that Topic Rank often fails to differentiate between high-level thematic definitions and critical results. As shown in Table 2. in Article 2. Topic Rank assigned high priority to a general definition of the process rather than the findings. Similar patterns were observed across multiple documents. indicating that these examples are representative of a systematic granularity mismatch between global thematic signals and specific research findings. While these sentences are relevant to the document's global topic. they lack the specific synthetic insights required for a concise abstract-level summary.

Table 2. Comparison between target abstracts and topic rank summaries

Article	Target Abstract	Topic Rank Summaries
1	*Penelitian ini menghasilkan sistem informasi yang dapat digunakan dalam proses pengusulan pembuatan kartu pegawai. kartu istri atau kartu suami pegawai…*	*Penelitian ini bertujuan untuk menghasilkan sistem informasi yang dapat digunakan dalam proses pengusulan pembuatan kartu…*
2	*Pelatihan merupakan usaha peningkatan kualitas sumber daya ditujukan untuk meningkatkan pengetahuan. keterampilan dan sikap setiap petugas atau pekerja yang berkaitan dengan kesehatan dan keselamatan kerja (K3) agar memiliki kompetensi sesuai…*	*Pelatihan ini meningkatkan kemampuan atau keterampilan terkait peningkatan keterampilan k3 melalui manajemen pelatihan dan kompetensi k3…*

The combination of TextRank. Keyword Count. and Position excels in unigram overlap (0.3390). yet configurations incorporating Semantic features demonstrate superior performance in meaning preservation. While the 5-feature combination remains robust at 0.356461. these findings challenge the common assumption in rank fusion that adding features always improves performance. particularly in scientific domains that are highly sensitive to semantic noise. The best configuration for lexical fidelity is identified as the 4-feature combination of TextRank. Keyword. Topic. and Semantic for unigram overlap (ROUGE-1: 0.3568). while the pairing of TextRank and Keyword is optimal for phrase-level precision (ROUGE-2: 0.1376).

The best configuration for semantic coherence and the most balanced configuration is the full 5-feature fusion (TextRank. Position. Keyword. Topic. and Semantic). which achieves a peak BERTScore of 0.6977 and preserves the structural flow of the IMRaD framework with a ROUGE-L of 0.2018. The convergence in performance between top models indicates that after reaching a specific threshold. improvements in summary quality depend more on the quality of feature representation and domain alignment than on the raw number of features integrated. This observation confirms that feature saturation occurs rapidly. adding redundant thematic features can introduce noise that degrades informational fidelity.

Discussion

Summary quality in the Indonesian scientific domain is governed by structural alignment rather than feature density. establishing that lexical overlap is an insufficient proxy for informational integrity. Our findings indicate that Indonesian scientific articles cannot be summarized effectively using a single-dimensional approach. instead there is no universal optimal configuration. but rather goal-dependent optimality. This feature saturation confirms that more features do not linearly improve performance. From a practical standpoint. the trade-off between ROUGE and BERTScore carries significant implications for human researchers. In the context of scientific synthesis. the accuracy of the meaning is far more critical than mere lexical similarity. A summary that uses the exact words of the original but fails to capture the causal relationship between variables is less useful to a researcher than a summary that might use synonyms or paraphrases but accurately conveys the findings. Therefore. configurations incorporating semantic features are objectively more useful for synthesis. as they prevent the misrepresentation of research results that can occur when relying solely on lexical fragments.

The identification of feature saturation reveals that it is a domain-dependent rather than a technical failure. In the Indonesian scientific domain. the rigid adherence to the IMRaD structure and the dense localization of technical terminology mean that increasing feature density beyond a certain threshold yield diminishing returns. This saturation is directly related to the structural rigidity of scientific articles. Once primary salience indicators. such as sentence position and keywords. are captured. additional thematic features often introduce redundant signals or semantic noise. Consequently. feature efficiency and strategic selection are more decisive for summary quality in scientific contexts than exhaustive feature integration.

The results demonstrate that the optimal configuration depends on whether the objective is lexical precision or semantic depth. The full 5-feature fusion which integrates TextRank. Position. Keyword. Topic. and Semantic ranks achieved a ROUGE-1 of 0.3564 and a BERTScore of 0.6977. When evaluating the impact of specific feature exclusions. the 4-feature configuration excluding Topic Rank (TextRank. Position. Keyword. and Semantic) yielded a ROUGE-2 of 0.1321 and a BERTScore of 0.6966. In contrast. the 4-feature configuration excluding Semantic Rank (TextRank. Position. Keyword. and Topic) resulted in a lower ROUGE-2 of 0.1287 and a BERTScore of 0.6955. This divergence confirms that the Semantic feature is instrumental in shifting the selection orientation from simple word matching to meaning matching. Semantic models prioritize core content and coherence over exact phrasing.

Consequently. Semantic Rank is more vital for informational integrity than Topic Rank. which often introduces a granularity mismatch. This is a limitation of LDA. where the model is optimized for document-level coherence over the specific research findings required for a concise abstract. while extractive summarization is a sentence-level decision process. This is a structural mismatch between document-level topic modeling and sentence-level extraction. While LDA seeks a global thematic balance. it effectively dilutes the salience of section-dependent results. leading to the selection of general background information rather than the specific synthetic insights necessary for scientific summarization.

The technical superiority of Keyword Count Rank over Topic Rank (LDA) at the sentence level is further explained by the concentration of technical terminology. In scientific discourse. high-value terms are not distributed uniformly. They are densely packed within specific IMRaD sections to describe experimental parameters. Keyword Count Rank acts as a high-precision local proxy because it identifies these pockets of information density. In contrast. LDA attempts to smooth the topic distribution across the entire document. By seeking a global thematic balance. LDA effectively dilutes the salience of section-dependent findings. often resulting in the selection of broad definitions rather than specific results. This granularity mismatch was evident in our failure analysis. where LDA prioritized background definitions over critical findings.

Unlike previous studies that assume generic lead-based extraction is sufficient. Indonesian scientific discourse follows rigid IMRaD-driven patterns. resulting in lead bias. The previous RankSum assumes that aggregating more features improves summary quality by mitigating the limitations of single-feature models (Joshi et al.. 2022). but our results demonstrate features saturation effect. where redundant thematic features yield diminishing returns and introduce semantic noise that degrades summaries' fidelity. While Fitrianah and Jauhari (2022) use paragraph location as a discrete positional feature based on sentence placement within paragraph. and Ulker and Ozer's (2024) model which summarizes abstractively by rewriting sentences. our results highlight that Position Rank must specifically favor the results and conclusion sections to correct biases toward background information. and maintains original sentences.

Our RankSum also corrects the existing Indonesian MMR systems (Faisal et al.. 2024). which rely on word-matching and lack sensitivity to the hierarchical IMRaD structure of scientific articles. The efficiency of the approach is technically realized through the synergy between TextRank and Position Rank. Although TextRank captures sentence centrality based on word connectivity. it tends to overselect background material from the Introduction. which Position Rank corrects by prioritizing structural importance to ensure the model selects sentences that are both logically central and aligned with the research's context.

Despite the effectiveness of the RankSum approach. its maximum potential is currently hampered by the limitations of the Indonesian NLP ecosystem. As a low-resource language in the scientific domain. Indonesian lacks large-scale contextual embedding models specifically trained on technical datasets. While models like IndoSBERT provide a baseline for semantic understanding. they may not fully capture the relationships between specialized Indonesian scientific terms. This provides a strong foundation for future research. which should prioritize the development of domain-specific embeddings like Indo-SciBERT. Furthermore. transitioning from LDA to neural topic models such as BERTopic or Top2Vec may resolve the granularity mismatch identified in this study. While rank fusion provides a robust framework. the Indonesian scientific domain demands a strategic selection of features that respect its structural rigidity and prioritize semantic integrity over lexical matching.

CONCLUSION

This study establishes three findings regarding the behavior of rank fusion in Indonesian scientific summarization. Rank fusion is inherently domain-sensitive. demonstrating that structural alignment with the IMRaD framework is more decisive than exhaustive feature integration. Empirical evidence confirms the phenomenon of feature saturation. where increasing the number of features yields diminishing returns and potentially introduces semantic noise. Semantic modeling is a necessity for scientific ATS to resolve the unavoidable trade-off between surface-level lexical fidelity and deep informational integrity. These findings suggest that for Indonesian scientific discourse. feature selection must prioritize categorical suitability over raw numerical quantity. The research is limited by the use of abstracts as reference summaries. the reliance on lexical weight optimization. and the inherent constraints of sentence-level extraction guided by document-level topic signals. Future work should address the broader challenges of low-resource scientific NLP by developing domain-specific

Indonesian embeddings. Incorporating factual consistency evaluation and human-in-the-loop summarization frameworks would further improve the qualitative reliability of automatic text summarization. Overall. domain-aware feature selection emerges as a more reliable foundation for scientific summarization in Indonesian NLP than architectural complexity or indiscriminate feature accumulation.

REFERENCES

Aurelia. M.. Monica. S.. & Girsang. A. S. (2024). Transformer-based abstractive indonesian text summarization. *International Journal of Informatics and Communication Technology (IJ-ICT). 13*(3). 388. https://doi.org/10.11591/ijict.v13i3.pp388-399

Azam. M.. Khalid. S.. Almutairi. S.. Ali Khattak. H.. Namoun. A.. Ali. A.. & Syed Muhammad Bilal. H. (2025). Current Trends and Advances in Extractive Text Summarization: A Comprehensive Review. *IEEE Access. 13*. 28150–28166. https://doi.org/10.1109/ACCESS.2025.3538886

Aziz. N. M. A.. Ali. A. A.. Naguib. S. M.. & Fayed. L. S. (2025). Clustering-based topic modeling for biomedical documents extractive text summarization. *The Journal of Supercomputing. 81*(1). 171. https://doi.org/10.1007/s11227-024-06640-6

Bandaru. R.. & Radhika. Y. (2022). Extractive multi-document text summarization leveraging hybrid semantic similarity measures. *International Journal of Advanced Computer Science and Applications. 13*(9). 844–852. https://doi.org/10.14569/IJACSA.2022.0130998

Cai. X.. Liu. S.. Yang. L.. Lu. Y.. Zhao. J.. Shen. D.. & Liu. T. (2022). COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers. *Journal of Biomedical Informatics. 127*. 103999. https://doi.org/10.1016/j.jbi.2022.103999

Faisal. M.. Mawaridi. B. H.. Afrah. A. S.. Supriyono. Arif. Y. M.. Aziz. A.. Wijayanti. L.. & Mulyadi. M. (2024). Enhancing Indonesian text summarization with latent dirichlet allocation and maximum marginal relevance. *International Journal of Advanced Computer Science and Applications. 15*(8). 519–528. https://doi.org/10.14569/IJACSA.2024.0150852

Fitrianah. D.. & Jauhari. R. N. (2022). Extractive text summarization for scientific journal articles using long short-term memory and gated recurrent units. *Bulletin of Electrical Engineering and Informatics. 11*(1). 150–157. https://doi.org/10.11591/eei.v11i1.3278

Girsang. A. S.. & Amadeus. F. J. (2023). Extractive text summarization for Indonesian news article using ant system algorithm. *Journal of Advances in Information Technology. 14*(2). 295–301. https://doi.org/10.12720/jait.14.2.295-301

Gulati. V.. Kumar. D.. Popescu. D. E.. & Hemanth. J. D. (2023). Extractive article summarization using integrated TextRank and BM25+ algorithm. *Electronics (Switzerland). 12*(2). https://doi.org/10.3390/electronics12020372

Joshi. A.. Fidalgo. E.. Alegre. E.. & Alaiz-Rodriguez. R. (2022). RankSum—An unsupervised extractive text summarization based on rank fusion. *Expert Systems with Applications. 200*(6). 116846. https://doi.org/10.1016/j.eswa.2022.116846

Joshi. A.. Fidalgo. E.. Alegre. E.. & Fernández-Robles. L. (2023). DeepSumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. Expert Systems with Applications. 211. https://doi.org/10.1016/j.eswa.2022.118442

Lau. A. J. J.. & Tan. C. W. (2024). LongT5Rank: A novel integrated hybrid approach for text summarisation. *Journal of Telecommunications and the Digital Economy. 12*(3). 73–96. https://doi.org/10.18080/jtde.v12n3.977

Liu. R.. Liu. M.. Yu. M.. Jiang. J.. Li. G.. Zhang. D.. Li. J.. Meng. X.. & Huang. W. (2024). GLIMMER: incorporating graph and lexical features in unsupervised multi-document summarization. *Frontiers in Artificial Intelligence and Applications*. *392*. 3709–3716. https://doi.org/10.3233/FAIA240930

Ma. T.. Pan. Q.. Rong. H.. Qian. Y.. Tian. Y.. & Al-Nabhan. N. (2022). T-BERTSum: topic-aware text summarization based on BERT. *IEEE Transactions on Computational Social Systems*. *9*(3). 879–890. https://doi.org/10.1109/TCSS.2021.3088506

Onan. A.. & Alhumyani. H. A. (2024). FuzzyTP-BERT: Enhancing extractive text summarization with fuzzy topic modeling and transformer networks. *Journal of King Saud University - Computer and Information Sciences*. *36*(6). 102080. https://doi.org/10.1016/j.jksuci.2024.102080

Ulker. M.. & Ozer. A. B. (2024). Abstractive summarization model for summarizing scientific article. *IEEE Access*. *12*. 91252–91262. https://doi.org/10.1109/ACCESS.2024.3420163

Wijaya. J.. & Girsang. A. S. (2024). Indonesian news extractive summarization using Lexrank and YAKE algorithm. *Statistics. Optimization and Information Computing*. *12*(6). 1973–1983. https://doi.org/10.19139/soic-2310-5070-1976

Yulianti. E.. Pangestu. N.. & Jiwanggi. M. A. (2023). Enhanced TextRank using weighted word embedding for text summarization. *International Journal of Electrical and Computer Engineering*. *13*(5). 5472–5482. https://doi.org/10.11591/ijece.v13i5.pp5472-5482

Zhang. X.. Wei. Q.. Song. Q.. & Zhang. P. (2024). TOMDS (Topic-Oriented Multi-Document Summarization): enabling personalized customization of multi-document summaries. *Applied Sciences*. *14*(5). 1880. https://doi.org/10.3390/app14051880

Zhou. Y.. Wei. J.. Sun. Y.. & Du. W. (2025). MP-UnSciBioSum: a multi-phase unsupervised document summarization method in scientific and biomedical domains. *Journal of King Saud University Computer and Information Sciences*. *37*(1–2). 11. https://doi.org/10.1007/s44443-025-00004-7