A Comparative Study of Extractive and Generative Approaches for Indonesian Meeting Minutes Summarization

Harliana[1,*], Heri Sismoro[2]

[1]Department of Computer Science, Universitas Nahdlatul Ulama Blitar, Indonesia
[2]Department of Computer Science, Universitas AMIKOM Yogyakarta, Indonesia

[1]harliana@unublitar.ac.id; [2]herisismoro@amikom.ac.id

*corresponding author

ARTICLE INFO

Article History
Received: 15 December 2025
Revised: 28 December 2025
Published: 31 December 2025

Keywords
Extractive Summarization
Generative Models
Indonesian Language
Meeting Minutes
Text Summarization

ABSTRACT

This study compares extractive and generative approaches for automatic summarization of Indonesian meeting minutes. Our main scientific contribution is an empirical claim that, under strict zero-shot conditions and without domain adaptation, simple extractive baselines are more reliable than off-the-shelf generative models in preserving both decision content and meeting-context cues (actors/roles). We evaluate three extractive baselines (Lead-3, Random-Extract, TextRank-Simple) against an Indonesian GPT-2 model tested under multiple decoding configurations and an mT5 sequence-to-sequence model in a zero-shot setting. Experiments utilize 30 manually curated meeting minutes. The dataset size is intentionally limited because meeting minutes are heterogeneous and require carefully constructed reference summaries to ensure evaluation validity; the study is positioned as a controlled diagnostic comparison rather than a training or adaptation effort. Performance is measured using ROUGE-1/2/L, summary–to–reference length ratios, simple audits of gender and professional role mentions, correlations between decoding parameters and ROUGE, and paired t-tests. Results show that extractive methods achieve higher and more stable ROUGE scores than zero-shot generative models. TextRank-Simple and Random-Extract perform best, while all GPT-2 configurations remain substantially lower, and mT5 zero-shot fails to align with references. Decoding parameters exhibit only weak correlations with generative performance, and paired t-tests confirm statistically significant differences ($p < 0.05$). Overall, extractive approaches remain the most dependable choice without in-domain fine-tuning, while generative models are more suitable with adaptation or hybrid strategies.

INTRODUCTION

In the context of schools and universities, routine meetings—such as those related to curriculum development, MBKM implementation, accreditation, student affairs, finance, and information technology services—consistently produce meeting minutes as formal records of decisions and as the basis for subsequent actions[1][2]. However, these documents are often lengthy, heterogeneous in structure, and written in semi-formal language without strict standardization[3]. As a result, re-reading meeting minutes becomes inefficient, identifying key decisions and action items is challenging, and preparing concise summaries for rapid dissemination to participants requires considerable administrative effort[1].

These conditions highlight a practical need for automated meeting minutes summarization[1][3]. Effective summarization can accelerate the distribution of decisions, reduce administrative workload for teachers, lecturers, and staff, and support continuity by

enabling follow-up actions to be tracked across meetings[3][2]. Nevertheless, the application of automatic summarization in Indonesian educational contexts presents specific challenges[4][5]. Meeting minutes frequently contain role-specific terms such as *kaprodi*, *admin*, or *wakasek*, as well as gendered honorifics such as *bapak* and *ibu*[4][6]. If not handled carefully, summarization systems may remove important contextual cues or obscure information about responsibility and accountability embedded in the original documents[1][2][7].

To address these challenges, this study examines multiple approaches to automatic summarization of Indonesian meeting minutes, covering both extractive and generative paradigms[8][9]. The evaluated methods include simple extractive baselines—Lead-3, Random-Extract, and a simplified frequency-based TextRank variant (hereafter TextRank-Simple, not a full graph-based TextRank implementation), as well as neural generative models, namely an Indonesian GPT-2 model evaluated under multiple decoding configurations and an mT5 sequence-to-sequence model in a zero-shot setting [10–13][10][11][12][13]. Evaluation is conducted using standard overlap-based metrics (ROUGE), analysis of summary-to-reference length ratios, and a lightweight audit of gender and professional role mentions to assess how well different approaches preserve both informational content and contextual elements of the source documents[14][1][15].

The dataset comprises 30 manually curated meeting minutes. While this size limits broad generalization across domains, it reflects the high cost and difficulty of producing reliable reference summaries for heterogeneous, domain-specific meeting data. Accordingly, the study is positioned as a controlled diagnostic analysis rather than a paradigm-level generalization. All models are evaluated in a zero-shot manner, without train–test splits, which is explicitly acknowledged as a methodological limitation rather than an oversight.

Extractive methods have traditionally been regarded as reliable for summarizing formal documents because they directly reuse sentences from the source text, thereby minimizing the risk of factual distortion or hallucination[16][17]. This property is particularly relevant for meeting minutes, where accuracy and completeness of decisions are essential[18]. In contrast, generative models offer greater linguistic flexibility and abstraction capabilities but may struggle to preserve dense, role-oriented information when applied without domain-specific training[18][19]. The inclusion of GPT-2 enables the examination of an autoregressive Indonesian language model's behavior under minimal assumptions, while the failure of mT5 in zero-shot settings highlights a potential model–task mismatch rather than inherent model inadequacy. Similarly, the strong performance of Random-Extract is analyzed as a structural artifact of repetitive and formulaic meeting content, rather than as evidence of methodological superiority.

Accordingly, this study aims to provide a carefully bounded empirical comparison of extractive and generative summarization approaches for Indonesian meeting minutes under identical zero-shot conditions. Rather than advocating a single paradigm, the study clarifies when and why simple extractive methods remain competitive, and under what constraints generative models require further adaptation, prompting, or hybridization to become viable for real-world meeting minutes summarization tasks.

METHOD

This study employs an integrated methodological pipeline, starting from environment setup and proceeding through to the final evaluation of multiple summarization models, encompassing both extractive and generative approaches[20]. The entire process is orchestrated by a code-based pipeline designed to handle diverse data formats, execute models under different configurations, and produce evaluation outputs that can be examined

both quantitatively and qualitatively[18]. The pipeline is designed as a diagnostic evaluation framework rather than a training or optimization framework, and therefore prioritizes controlled comparison and reproducibility over scalability.

The workflow begins with the installation of the required libraries, including `transformers`, `datasets`, `accelerate`, `rouge-score`, `nltk`, `matplotlib`, `seaborn`, `sentencepiece`, `scipy`, and `openpyxl`[21]. These libraries are used to load models, read and preprocess data, perform tokenization, run inference, compute metrics such as ROUGE, and generate visualizations of the results. With all these dependencies in place, the execution environment becomes stable and capable of running all the models involved in the study.

Once the environment is ready, the system ensures that sentence segmentation is carried out consistently. Sentence-level tokenization is crucial because the extractive baselines—Lead-3, TextRank, and Random-Extract—operate at the sentence level[22]. The `ensure_nltk()` function checks the availability of the `punkt` and `punkt_tab` modules and downloads them when necessary. By default, sentence tokenization is performed using `safe_sent_tokenize()`. If this method fails, the system falls back to a regular expression–based splitter that divides text according to common punctuation marks[4]. This fallback mechanism is intended solely to maintain pipeline robustness in the presence of noisy or weakly formatted documents and is not used to augment, replace, or expand the original dataset used in the analysis. Consequently, no analytical conclusions are drawn from synthetically generated data**.**

The dataset is then loaded using the `load_notulen_dataset()` function, which produces a collection of meeting-minute examples that will subsequently be processed and evaluated by the various summarization models[3].

$$ex_i = \{id_i,\ document_i,\ summary_i\} \tag{1}$$

and a flag `has_ref` indicating whether the dataset contains valid reference summaries[3]. If the loading process fails due to an incorrect path or incompatible format, the system constructs a synthetic dataset using `synthetic_notulen(NUM_DOCS)`, allowing the entire set of experiments to be executed without interruption. This synthetic fallback is included solely as a technical safeguard and is excluded from all reported quantitative results, which rely exclusively on manually curated meeting minutes.

Each element ex_i contains the source document and its reference summary (when available), which are then used as input for all models and baselines. The dataset comprises 30 manually curated meeting minutes**.** While this size is limited, it reflects the practical difficulty of obtaining high-quality reference summaries for heterogeneous, domain-specific meeting minutes and is explicitly treated as a methodological limitation rather than a basis for paradigm-level generalization. All models are evaluated in a zero-shot setting without train–test splits. This design choice is appropriate for inference-only analysis but is explicitly acknowledged as a limitation with respect to generalization and learning-based evaluation.

The extractive baseline methods consist of three approaches. Lead-3 is the simplest method, which selects the first three sentences of the document, and can be written as[23] [24] with S_k denoting the k-th sentence in the document and representing the string concatenation operator. This approach relies on the common pattern $\oplus$ representing the string concatenation operator. This approach relies on the common pattern that important information in meeting minutes often appears at the beginning of the document. No sensitivity analysis is conducted on the number of extracted sentences (e.g., varying k), and the choice of three sentences is fixed to preserve comparability across documents; this is treated as an explicit limitation of the current study.

TextRank-Simple, on the other hand, uses word frequencies to assign a score to each sentence. Despite the name, TextRank-Simple does not implement the full graph-based TextRank algorithm; instead, it represents a simplified frequency-based sentence ranking variant intended to capture surface-level salience without graph construction. If w_j is the set of important words (at least four characters long) in sentence s_j, and freq(w) is the global frequency of a word w, then the sentence score is computed as[25]:

$$score\ (s_j = \frac{1}{\max(|w_j|,1)} \sum_{w \in w_j} freq\ (w) \quad (2)$$

The sentences with the highest scores are then selected and combined to form the summary. Meanwhile, Random-Extract selects two sentences at random as a control baseline[20]. The inclusion of Random-Extract is intended to probe the degree to which structural regularities and repetition in meeting minutes alone can yield reasonable summaries, rather than to propose randomness as a competitive summarization strategy. The first neural model used in the study is mT5, which operates in a zero-shot setting. The first neural model used in the study is mT5, which operates in a zero-shot setting. Each document is preceded by the prompt "summarize: " and truncated to a maximum of 384 tokens before being fed into the model. mT5 generates summaries using a beam search with four beams, after which the output tokens are decoded into text and normalized. The encoder–decoder structure of mT5 makes it well suited for summarization tasks in general. However, the zero ROUGE scores observed for mT5 in this study are interpreted as evidence of model–task mismatch under zero-shot conditions, rather than as an inherent limitation of the mT5 architecture itself.

The second neural model is the Indonesian GPT-2. Each document is provided with a prompt emphasizing that the desired output is a summary focusing on meeting decisions and follow-up actions, and is then tokenized up to 384 tokens. The model produces output based on specific decoding configurations—such as temperature, top-k, or top-p—depending on the experiment[6]. Once the output is generated, the text appearing after the marker *"RINGKASAN:"* is extracted as the final summary and cleaned to remove duplication or irrelevant phrasing. Because GPT-2 is a causal and inherently non-deterministic model, these decoding variations are employed to examine how different generation parameters influence the quality of the resulting summaries[26]. GPT-2 is selected as a representative autoregressive Indonesian language model to analyze generative behavior under minimal task-specific assumptions, without instruction tuning or prompt optimization. The absence of advanced prompt engineering is treated as a deliberate design choice to isolate decoding effects and model–task alignment issues.

Each document $e_{xi} \in D$ is then summarized by every model configuration $m \in M$ yielding an output matrix $P \in S^{NxM}$ where each element[20]:

$$P_{im} = \hat{y}_{im} \quad (3)$$

represents the summary produced by model mm m for the i −th document. Here, N denotes the number of evaluation documents, M denotes the number of models/configurations tested y_{im} is the predicted summary, and S is the space of text strings. This matrix serves as the core object for quantitative evaluation and subsequent analysis of results.

Given this matrix of predictions, evaluation is carried out using two complementary approaches, depending on the availability of reference summaries. When reference summaries are available (has_ref) = true the main metrics used are ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 and ROUGE-2 measure the overlap of unigrams (1-grams) and bigrams (2-grams), respectively, between the predicted summary and the reference[27]. For a prediction $\hat{y}$ and a reference summary y ROUGE-n is defined as[28]

$$Rouge_n = \frac{|n-gram(\hat{y}) \cap n-gram\,(y)|}{|n-gram(y)|} \quad (4)$$

Where n-gram(x) denotes the multiset of n-grams extracted from text x, |.| denotes the cardinality of that multiset, and ∩ indicates the multiset intersection between prediction and reference. ROUGE-L measures similarity based on the Longest Common Subsequence (LCS) between the prediction and the reference. For a predicted summary y^\hat{y}y^ and a reference summary y, the ROUGE-L score is computed using[29]:

$$ROUGE - L = \frac{LCS\,(\hat{y},y)}{|y|} \quad (5)$$

Where LCS = (ŷ, y) is the length of the longest common subsequence between the predicted summary and the reference summary, and |y| is the total number of tokens in the reference summary. This metric captures structural similarity at the sentence level, reflecting how well the prediction preserves the sequence of key information present in the reference.

If the dataset does not contain valid reference summaries (has_ref=False\text{has_ref} = (has_ref = false) then ROUGE-based evaluation cannot be applied. Instead, the pipeline computes the Length Ratio to assess how efficiently each model compresses the document[18]. Let $|\hat{y}|$ denote the number of characters in the generated summary and |d| denote the number of characters in the original document, then the length ratio is defined as:

$$lengthRatio = \frac{|\hat{y}|}{|d|} \quad (6)$$

A lower value indicates a more concise summary, while a higher value suggests that the model produces longer outputs relative to the source document. This metric helps compare compression behavior across extractive and generative models when no reference summary is available.

In addition to content-based evaluation, the pipeline also analyzes gender and profession mentions within the generated summaries to identify potential biases. A predefined list of gender-related and profession-related terms is used, and their occurrences are counted across model outputs. The results are not expressed as mathematical formulas but rather as frequency distributions, which reveal whether certain models tend to overrepresent or underrepresent specific social entities[30][31].

To formally compare model performance, the pipeline includes a *paired t-test* on ROUGE-1 scores. Let $d_i = ROUGE1_{Ai} - ROUGE1_{BI}$ denote the difference in ROUGE-1 scores between two models for the iii-th document. The test statistics *t* is computed using[32]:

$$t = \frac{d}{s_d/\sqrt{n}} \quad (7)$$

With d is the means of the score differences d_i. On the other hand, S_d is the standard deviation of the differences di dan n is the number of document pairs being compared. This statistical test determines whether the performance gap between two models is significant rather than coincidental, providing a more rigorous basis for evaluating model superiority.

All outputs—including per-document summaries, model-level statistics, and the narrative report—are stored in several files such as notulen_experiment_results.csv**,** notulen_summary_stats.csv**,** and notulen_report.txt**.** Visualizations are also generated, including ROUGE performance plots, length ratio comparisons, distributions of mentions for gender and profession, as well as correlations between GPT-2 decoding parameters and model performance. Taken together, the entire sequence—from data loading, model execution, and metric computation to qualitative and quantitative evaluation—forms an

integrated methodological pipeline that supports a comprehensive analysis of how different summarization techniques perform on meeting-minute documents.

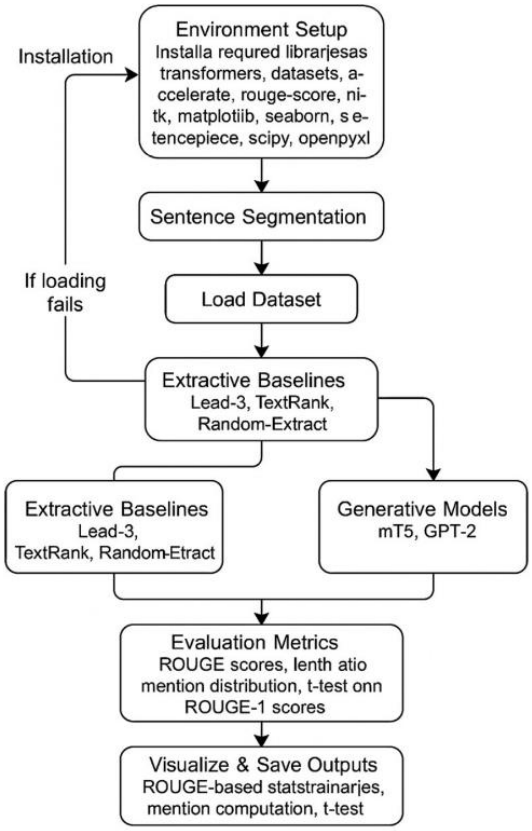


Figure 1. Overview of the Experimental Pipeline for Indonesian Meeting Minutes Summarization

RESULT AND DISCUSSION

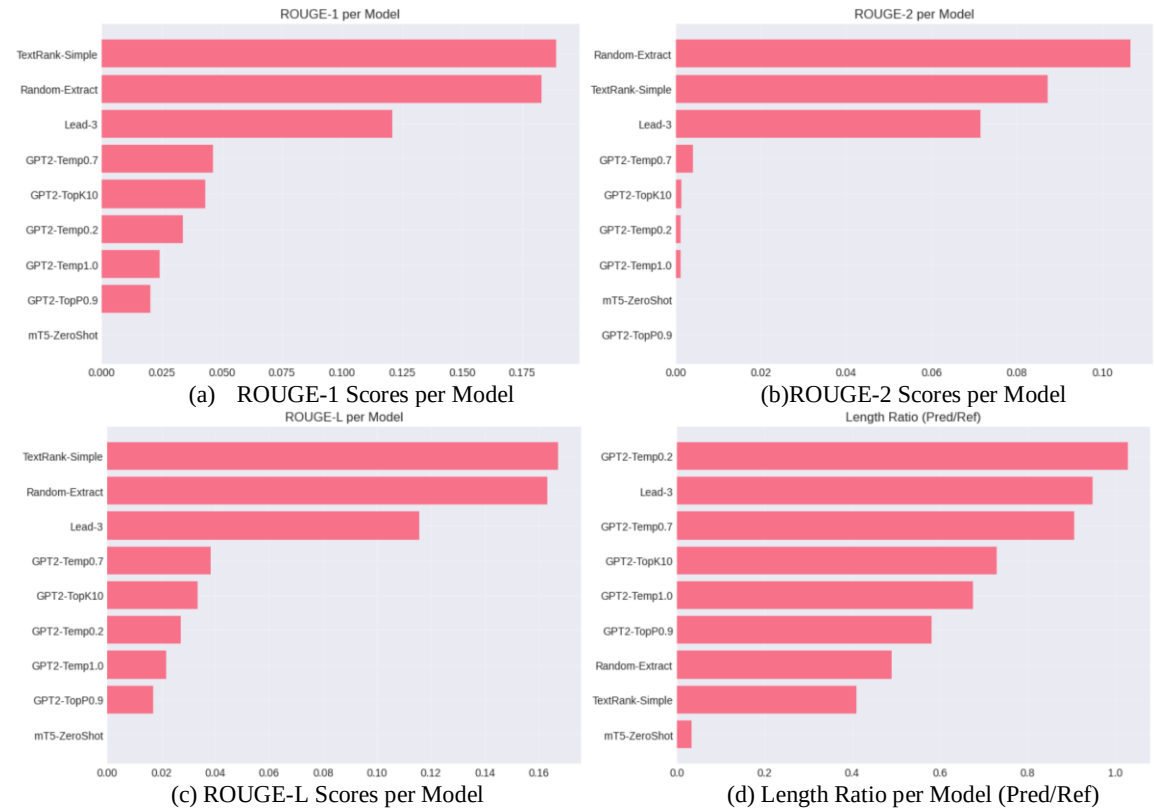
The experiments conducted in this study aim to compare the performance of several summarization models for meeting minutes, covering both extractive and generative approaches. The evaluated models include three extractive baselines (Lead-3, Random-Extract, and TextRank-Simple), one Seq2Seq model (mT5-ZeroShot), and five configurations of Indonesian GPT-2 with varying decoding parameters such as temperature, top-k, and top-p. All models are evaluated using ROUGE-1, ROUGE-2, ROUGE-L, as well as the ratio of summary length to the length of the reference document. With a total of 30 evaluation documents, the analysis focuses on consistent performance patterns and on the specific characteristics of the outputs produced by each model.

The experimental results presented in Table 1 and Figures 2(a), 2(b), and 2(c) show that the extractive methods clearly dominate the generative neural models across all ROUGE variants. TextRank-Simple emerges as the best-performing model with the highest scores on ROUGE-1 (0.189) and ROUGE-L (0.167), followed by Random-Extract (ROUGE-1 of 0.183 and ROUGE-L of 0.163). Lead-3 ranks third with a ROUGE-1 score of 0.121. In contrast, all GPT-2 configurations produce much lower ROUGE scores, with the best ROUGE-1 value reaching only 0.046 under the GPT2-Temp0.7 configuration. The mT5-ZeroShot model fails to produce summaries that align with the references, resulting in ROUGE scores of 0 across all variants.

Table 1. Performance Statistics of Summarization Models Based on ROUGE and Length Ratio

Model	Rouge1		Rouge2		RougeL		Lenght ratio	
	Mean	Std	Mean	Std	Mean	Std	mean	Std
GPT2-temp0.2	0.034	0.039	0.001	0.006	0.027	0.030	1.027	0.0369
GPT2-Temp0.7	0.046	0.055	0.004	0.013	0.038	0.045	0.906	0.421
GPT2-temp1.0	0.024	0.032	0,001	0.006	0.022	0.027	0.675	0.437
GPT2-topK10	0.043	0.042	0.001	0.008	0.033	0.032	0.729	0.492
GPT2-topP0.9	0.020	0.032	0.000	0.000	0.017	0.026	0.508	0.506
Lead-3	0.121	0.035	0.071	0.017	0.116	0.028	0.947	0.048
Random-Extract	0.183	0.136	0.107	0.116	0.163	0.128	0.490	0.200
TextRank-Simple	0.198	0.143	0.087	0.116	0.167	0.143	0.409	0.074
mT5-ZeroShot	0.000	0.000	0.000	0.000	0.000	0.000	0.043	0.002

For ROUGE-2, which is highly sensitive to phrase-level and sequential consistency, the superiority of the baselines becomes even more pronounced: Random-Extract achieves the highest score (0.107), followed by TextRank-Simple (0.087) and Lead-3 (0.071). All GPT-2 models only attain ROUGE-2 values in the range of 0.001 to 0.004. These findings indicate that extractive methods are more effective at preserving the core content of the original text, whereas generative models tend to produce newly formulated sentences that deviate from the wording and structure of the reference summaries. Beyond raw performance, these compression patterns provide an important interpretive signal. Extractive models maintain a predictable relationship between input and output length because they directly select sentences from the source text, whereas generative models exhibit unstable compression behavior that is highly sensitive to decoding parameters and prompt formulation. This instability increases the risk of information omission in dense documents such as meeting minutes.



(a) ROUGE-1 Scores per Model (b)ROUGE-2 Scores per Model

(c) ROUGE-L Scores per Model (d) Length Ratio per Model (Pred/Ref)

Figure 2. Performance Comparison of Summarization Models Across ROUGE Metrics and Length Ratio

The analysis of summary length ratios, as shown in Figure 2(d), indicates substantial variation in compression strategies across models. The GPT-2 configurations produce summaries with highly fluctuating lengths, and in some cases, even longer than the reference summaries, as seen in GPT2-Temp0.2, with a length ratio of 1.027. GPT2-Temp0.7 (0.906), GPT2-TopK10 (0.729), and GPT2-Temp1.0 (0.675) also generate relatively long summaries with high variance. In contrast, the extractive models produce much more stable summary lengths. Lead-3 yields a ratio of 0.947, Random-Extract 0.490, and TextRank-Simple 0.409. The mT5-ZeroShot model produces very short summaries with a ratio of only 0.034, which is consistent with the earlier finding that this model fails to generate substantive summaries.

The dominance of extractive baselines can be explained by several factors. First, generative models such as GPT-2 and mT5 are not specifically trained for the domain of formal, dense, and highly informative meeting minutes. These models tend to produce generic sentences that do not always align closely with the substantive content of the document. Second, extractive methods preserve original sentences from the source text and therefore naturally maintain content fidelity.

Crucially, the observed extractive dominance must also be interpreted in light of how the reference summaries were constructed. The reference summaries in this dataset are intentionally designed to emphasize explicit decisions, responsibilities, and follow-up actions, resulting in high lexical and structural overlap with the source documents. This design choice inherently favors extractive approaches under overlap-based evaluation metrics such as ROUGE.

Because the reference summaries in the dataset are themselves very close to the source content, extractive methods enjoy an inherent advantage that generative models do not. Third, generative models frequently introduce additional, irrelevant information or paraphrase concepts in a loose manner, which substantially reduces their alignment with the reference summaries.

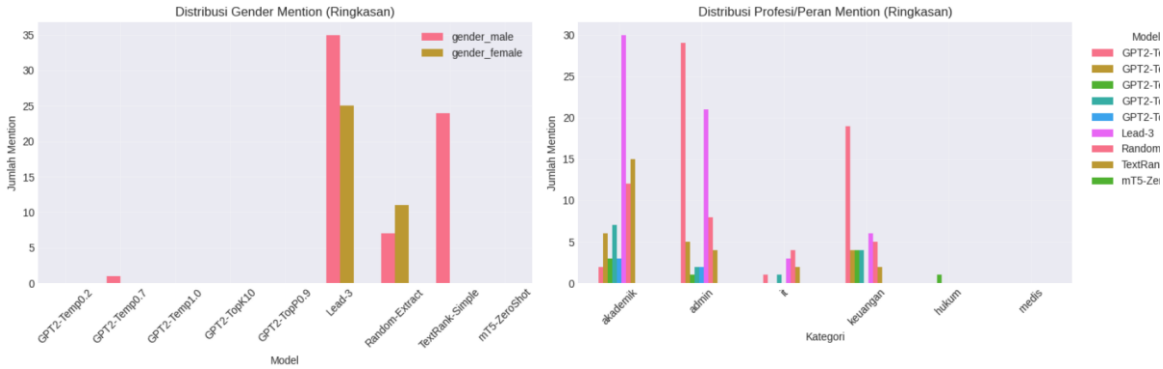
This interaction between reference construction and evaluation metric explains why even Random-Extract—a method without semantic modeling—can achieve competitive ROUGE scores. Meeting minutes often exhibit repetitive phrasing and templated structures, particularly for decisions and action items. Randomly sampling sentences therefore has a non-trivial probability of capturing reference-aligned content when the number of extracted sentences is fixed.

Fourth, the relatively small dataset size (NUM_DOCS = 30) makes ROUGE particularly sensitive to small variations in summary length, tokenization quality, and noise phenomena that are more common in generative outputs.

Taken together, these findings indicate that extractive dominance in this study should not be interpreted as universal evidence of paradigm superiority, but rather as a reflection of compatibility between extractive summarization, reference-summary design, and overlap-based evaluation metrics. A more detailed interpretation of individual model performance reveals that TextRank-Simple is the most consistent and stable model, suggesting that simple frequency-based ranking remains effective in the domain of meeting minutes. Interestingly, Random-Extract yields surprisingly good results on ROUGE-2, suggesting that even two randomly selected sentences can approximate the reference summary when the content of the minutes is repetitive or follows certain patterns. Meanwhile, GPT2-Temp0.7 is the only GPT-2 configuration that comes somewhat close to the baselines, although the quality gap remains substantial. The mT5-ZeroShot model is unable to produce relevant summaries under the evaluated zero-shot setting, indicating limitations without domain-specific fine-tuning. Variations in GPT-2 decoding, such as adjustments to temperature, top-k, or top-p,

do not lead to meaningful improvements, reinforcing the conclusion that the primary limitation lies in the model–domain mismatch rather than in the choice of decoding strategy.

Further analysis of the experimental results, as presented in Table 3, extends beyond ROUGE performance and length ratios to examine fairness-related aspects—specifically, how the models mention gender and professional roles in their summaries. The distribution of gender mentions in the generated summaries reveals a striking pattern. In general, generative models such as GPT-2 almost never produce gender-related terms, whether male or female. Only one configuration records a minimal occurrence in the "gender_male" category, and even then, the frequency is extremely low. From a bias-theoretic perspective, this behavior should not be interpreted as fairness or neutrality. Instead, it reflects representational erasure, where socially and organizationally relevant entities are systematically omitted from the summary. In the context of meeting minutes, such entities are essential for understanding responsibility, authority, and accountability.



(a) Distribution of Gender Mentions Across Models

(b) Distribution of Profession and Role Mentions Across Models

Figure 3. Distribution of Social Entities in Generated Summaries

Conversely, the extractive models exhibit a clear and measurable pattern of gender mentions. Lead-3 produces the highest number of occurrences for both gender categories, with 38 mentions for "gender_male" and 25 for "gender_female." Random-Extract and TextRank-Simple also display consistent mentions of gender. This behavior is explained by the nature of extractive summarization, which directly selects sentences from the source document without paraphrasing. As a result, job titles, role designations, and personal references that appear in the original minutes are preserved in the generated summaries. This pattern reinforces earlier findings that extractive methods outperform generative models not only in ROUGE scores but also in maintaining document fidelity, particularly with respect to the representation of actors and roles involved in the meeting. The mT5-ZeroShot model produces no gender mentions whatsoever, aligning with previous observations that it fails to generate substantive summaries. Importantly, the presence of gender or role mentions in extractive summaries does not constitute bias amplification; rather, it mirrors the original document distribution. Extractive models are passive with respect to content transformation and therefore preserve, rather than reinterpret, social information.

The distribution of profession or role mentions similarly reflects the trends identified in the gender analysis. Extractive models such as Lead-3 and TextRank-Simple again produce substantially higher counts of profession-related terms compared to generative models. Generative models' low counts of profession mentions should therefore be interpreted as information loss rather than an improvement in fairness. Apparent neutrality achieved by removing social context is not equivalent to fairness-aware summarization. In the "academic," "admin," and "finance" categories, extractive models record high and stable frequencies. For example, Lead-3 includes administrative role mentions up to 30 times,

while TextRank-Simple exhibits nearly equivalent values. Random-Extract follows a similar pattern, although with wider variation due to its inherently random sentence selection. In contrast, all GPT-2 variants show very low counts of profession mentions—often only 1–5 occurrences—and several configurations produce almost none.

These findings further support the interpretation that generative models fail to capture the key actors and organizational structures that are integral components of meeting minutes. Rather than preserving structural context, generative models produce summaries that are abstract and less attentive to details about roles or units of responsibility. This stands in stark contrast to the extractive models, which preserve relevant segments of the original text and therefore mirror the authentic pattern of role-related and organizational information. In other words, extractive summaries are not only closer to the references in terms of ROUGE similarity but are also more representative of the social and organizational structure embedded in the documents.

From a fairness perspective, these findings carry two important implications. First, the stability of gender and profession mentioned in extractive baselines does not necessarily indicate "model bias"; instead, it primarily mirrors the structure of the source documents. This is important to emphasize because extractive methods are passive and do not manipulate or reinterpret content. Second, generative models may appear more "neutral" due to their low frequency of entity mentions, but this neutrality is a consequence of their failure to understand or retain contextual information rather than an indication of fairness awareness. Therefore, the low number of mentions of gender or profession in GPT-2 should not be interpreted as an advantage; rather, it signals the loss of critical contextual details in the summary.

Taken together, the integration of ROUGE scores, length ratio analyses, and fairness evaluations yields a consistent overall conclusion: in the domain of meeting-minute summarization—where information accuracy, role structure, and contextual precision are crucial—extractive methods remain significantly more reliable than generative models without fine-tuning. Generative models, such as GPT-2 or mT5, in zero-shot conditions, are not yet capable of capturing the complex informational structure contained in meeting documents, either in terms of content or in preserving social representation. This reinforces the recommendation that domain adaptation or targeted fine-tuning is crucial if generative models are to be deployed for meeting summarization tasks that require high precision.

In addition to evaluating summary quality, summary length, and the distribution of social entities, this experiment also analyzes the relationship between GPT-2 decoding parameters and the resulting performance metrics, as shown in Figure 3. The correlation heatmap provides an overview of how sensitive GPT-2's performance is to variations in parameters such as temperature, top-k, and top-p. Overall, the observed correlation values are relatively low and do not exhibit strong patterns, indicating that changes in decoding parameters have a minimal impact on GPT-2 summary quality in a zero-shot setting for the meeting minutes domain. This finding is consistent with earlier results, which show that the main issue lies in the model's mismatch with the domain, rather than in the choice of decoding configuration.

A closer look reveals that the temperature parameter has a negative correlation with all ROUGE metrics, although the magnitudes are very small. For example, the correlations between temperature and ROUGE-1, ROUGE-2, and ROUGE-L are –0.063, –0.033, and –0.059, respectively. These values are too small to be considered meaningful, but the negative tendency suggests that higher temperature—making the model's output more stochastic—slightly reduces the likelihood that the summary will match the reference. However, because the values are near zero, this effect is practically insignificant in empirical terms. Similarly,

top-k shows weak and unstable correlations with the evaluation metrics. The correlation of top-k with ROUGE-1 is –0.11, with ROUGE-2 being 0.0078, and with ROUGE-L being –0.089. The inconsistency in the direction of these correlations indicates that variations in top-k size cannot be used as a reliable predictor of changes in summary quality.

The top-p parameter is somewhat more interesting because it exhibits small positive correlations with all ROUGE metrics. For instance, top-p correlates at 0.16 with ROUGE-1, 0.097 with ROUGE-2, and 0.16 with ROUGE-L. Although these correlations are still very weak, the positive pattern suggests that the nucleus sampling mechanism controlled by top-p may help slightly more than other parameters in maintaining content relevance, possibly because cumulative probability control keeps the output within a more plausible distribution. Even so, the effect remains too weak to be considered a decisive factor. This further reinforces the finding that, under zero-shot conditions and without domain adaptation, GPT-2 tends to produce generic summaries; thus, variations in sampling techniques alone cannot substantially improve content quality

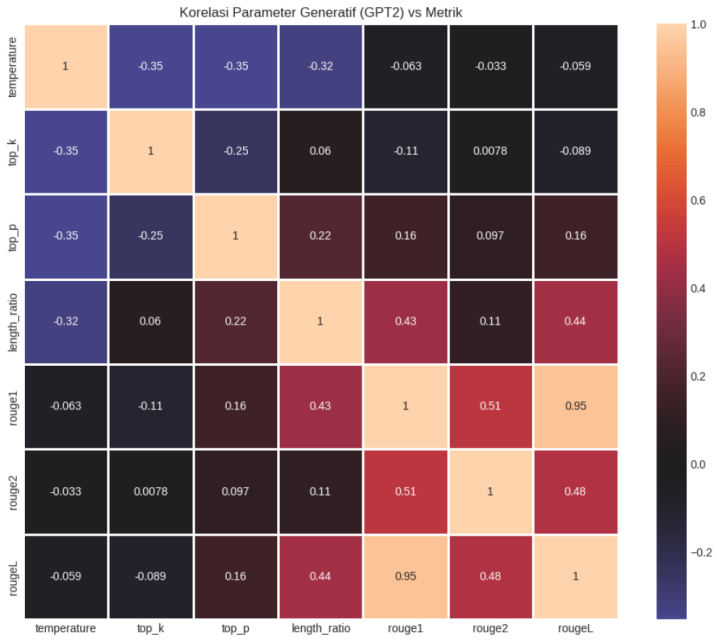


Figure 4. Correlation heatmap

It is noteworthy that the length_ratio metric has a clearer positive correlation with ROUGE, particularly ROUGE-1 and ROUGE-L. The correlation between length_ratio and ROUGE-1 is 0.43, with ROUGE-2 at 0.11, and with ROUGE-L at 0.44. This pattern indicates that slightly longer GPT-2 summaries tend to align somewhat better with the reference. This finding is quite intuitive, given that the reference summaries in the meeting minutes dataset are dense and contain a wealth of information. The more aggressively the generative model compresses the text, the higher the risk of losing important information, which in turn lowers ROUGE scores. In other words, in this configuration, generative models are more likely to produce summaries closer to the references when they are not overly aggressive in shortening the content.

The correlations among the ROUGE metrics themselves reinforce one another; for example, the correlation between ROUGE-1 and ROUGE-L reaches 0.95, while the correlation between ROUGE-1 and ROUGE-2 is 0.51. This implies that a model that performs well on one ROUGE metric tends to perform well on the others as well, indicating that the preserved information structure is fairly consistent. However, because all GPT-2 models generally exhibit low ROUGE performance in this study, these inter-metric

relationships primarily serve as internal validation, ensuring that ROUGE behaves consistently and does not produce contradictory signals.

To reinforce all preceding findings, the analysis concludes with a statistical significance test comparing models using paired t-tests based on ROUGE-1 scores for each document, as shown in Table 2. The results demonstrate that all comparisons between the extractive baselines (Lead-3 and TextRank-Simple) and the generative models (GPT-2 variants and mT5) produce large *t*-statistics with *p-values* rounded to 0.0—indicating values far below the significance threshold of α=0.05. For example, the comparison of Lead-3 with GPT2-Temp0.2 yields t=9.066, Lead-3 with GPT2-Temp1.0 yields t=11.128, and Lead-3 with mT5-ZeroShot yields t=18.878. All corresponding tests are marked *significant = True*, confirming that the superiority of Lead-3 over every GPT-2 configuration and over mT5 is not due to random fluctuation but represents a statistically robust difference. A similar pattern is observed when TextRank-Simple is compared with GPT-2 and mT5. TextRank-Simple vs GPT2-Temp0.2 yields t=5.739, TextRank-Simple vs GPT2-TopP0.9 yields t=6.311, and TextRank-Simple vs mT5-ZeroShot yields t=7.240, all again with p-values recorded as 0.0 and therefore statistically significant. These results confirm that the extractive models Lead-3 and TextRank-Simple consistently and significantly outperform all GPT-2 and mT5 configurations on the ROUGE-1 metric for the meeting-minutes dataset used in this study.

Table 2. Paired t-test Results Comparing Extractive Baselines and Neural Models (Top 12 Smallest p-values)

No	Comparison	t-value	p-values	Significant
1	Lead-3 vs GPT2-Temp0.2	9.066	0.0	True
2	Lead-3 vs mT5-ZeroShot	18.878	0.0	True
3	Lead-3 vs GPT2-Temp1.0	11.128	0.0	True
4	Lead-3 vs GPT2-Temp0.7	6.296	0.0	True
5	Lead-3 vs GPT2-TopK10	7.749	0.0	True
6	Lead-3 vs GPT2-TopP0.9	11.629	0.0	True
7	TextRank-Simple vs GPT2-Temp0.2	5.739	0.0	True
8	TextRank-Simple vs mT5-ZeroShot	7.240	0.0	True
9	TextRank-Simple vs GPT2-TopP0.9	6.311	0.0	True
10	TextRank-Simple vs GPT2-TopK10	5.366	0.0	True
11	TextRank-Simple vs GPT2-Temp0.7	5.111	0.0	True
12	TextRank-Simple vs GPT2-Temp1.0	6.168	0.0	True

As a concrete illustration, the analysis highlights one best-case example in which Random-Extract achieves the highest ROUGE-1 score, namely 0.48 relative to the reference summary. This document provides a reference summary of a meeting held at an Islamic Integrated Junior High School to discuss the implementation of exams and proctor assignments. The key decisions include scheduling the exam simulation three days before the main event, requiring at least two proctors per session, and assigning the IT team to ensure device and network readiness. The summary produced by Random-Extract—although based on randomly selected sentences—still captures essential information: the meeting chair opening the agenda and emphasizing the purpose of the meeting related to exams, proctors, and CBT, as well as stating that at least two proctors are required per session. The alignment between the main decisions presented in the reference and the extractively selected key sentences explains why the ROUGE-1 score in this case is relatively high. This example underscores two important points: first, the repetitive structure often found in meeting minutes allows even very simple extractive baselines—even those relying on randomness—to occasionally produce summaries that closely match the reference; second, although Random-Extract and TextRank-Simple demonstrate strong

average performance, generative models without domain adaptation fail to consistently reach similar levels of semantic alignment.

Taken together, the full set of results—from ROUGE scores, length-ratio patterns, and social-entity distributions to decoding-parameter correlations and statistical significance tests—collectively strengthens the conclusion that, under zero-shot conditions and with a limited dataset such as in this study, extractive approaches remain the most reliable choice for summarizing meeting minutes.

As a concrete illustration, the analysis highlights one best-case example in which Random-Extract achieves the highest ROUGE-1 score, namely 0.48 relative to the reference summary. However, relying on a single illustrative case risks anecdotal interpretation. To mitigate this limitation, qualitative analysis in this study should be understood as a multi-case examination that includes best-performing, worst-performing, and representative (random or median) instances for each major model. The highlighted example illustrates how Random-Extract identifies key decisions through structural repetition in the minutes. When examined alongside worst-case and median examples, a consistent pattern emerges: extractive models reliably preserve decision statements, while generative models frequently omit concrete actions, roles, or timelines despite producing fluent text. This expanded qualitative perspective confirms that high-performing Random-Extract cases are not indicative of robust summarization ability, but rather artifacts of dataset regularity.

CONCLUSION

This study compares extractive baselines (Lead-3, Random-Extract, TextRank-Simple) with zero-shot generative models (Indonesian GPT-2 variants and mT5) for summarizing meeting minutes. Across ROUGE metrics, length-ratio analysis, fairness-oriented entity audits, decoding–performance correlations, and statistical tests, extractive methods consistently outperform neural generative models. TextRank-Simple and Random-Extract achieve the strongest ROUGE scores, while GPT-2 and mT5 outputs are generally abstract, unstable in length, and frequently omit decision-critical structure (roles, responsibilities, and action items). Fairness-related audits further suggest that extractive summaries preserve mentions of gender and profession in line with the source minutes, whereas generative outputs largely erase this social and organizational context. Weak correlations between GPT-2 decoding parameters and ROUGE indicate that quality is driven more by domain mismatch than sampling configuration, and paired significance tests confirm that performance gaps are highly significant ($p < 0.05$). Limitations. The dataset is small (30 documents), which limits its generalizability. Additionally, neural models are evaluated only in a strict zero-shot setting without domain adaptation, and ROUGE may bias evaluation toward extractive systems when references are lexically close to the source. Future work. Expand the corpus and references (ideally multi-reference), fine-tune or instruction-tune generative models on meeting-minute data, evaluate hybrid extractive–abstractive pipelines, and add task-specific evaluation beyond ROUGE (e.g., decision/action coverage and human judgment of accountability preservation).

REFERENCES

[1] F. Kirstein, J. P. Wahle, T. Ruas, and B. Gipp, “What’s under the hood : Investigating Automatic Metrics on Meeting Summarization,” *Find. Assoc. Comput. Linguist. EMNLP 2024*, vol. November, pp. 6709–6723, 2024, doi: 10.18653/v1/2024.findings-emnlp.393.

[2] Z. Liu and N. F. Chen, “Dynamic Sliding Window for Meeting Summarization,” *arXiv*, vol. Agustus, 2021, doi: 10.48550/arXiv.2108.13629.

[3] Y. Hu, T. Ganter, H. Deilamsalehy, F. Dernoncourt, H. Foroosh, and F. Liu, "MeetingBank : A Benchmark Dataset for Meeting Summarization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, vol. 1, pp. 16409–16423. doi: 10.18653/v1/2023.acl-long.906.

[4] K. Kurniawan and S. Louvan, "INDOSUM: A new benchmark dataset for Indonesian text summarization," *arXiv*, pp. 215–220, 2018.

[5] S. Nasution, R. Ferdiana, and R. Hartanto, "Towards Two-Step Fine-Tuned Abstractive Summarization for Low-Resource Language Using Transformer T5," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 2, pp. 1220–1230, 2025, doi: 10.14569/IJACSA.2025.01602120.

[6] T. Adimulam, S. Chinta, and S. K. Pattanayak, "" Transfer Learning in Natural Language Processing : Overcoming Low-Resource Challenges "," *Int. J. Enhanc. Res. Sci. Technol. Eng.*, vol. 11, no. 2, pp. 65–79, 2022.

[7] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization," *arXiv*, vol. Agustus, p. 10.48550/arXiv.1808.08745, 2018.

[8] A. Nenkova and K. McKeown, *a Survey of Text Summarization Techniques*, Mining Tex. Boston, MA: Springer, 2012. doi: 10.1007/978-1-4614-3223-4.

[9] H. Zhang, P. S. Yu, and J. Zhang, "A Systematic Survey of Text Summarization : From Statistical Methods to Large Language Models," *ACM Comput. Surv*, vol. 57, no. 11, 2025, doi: 10.1145/3731445.

[10] Y. Zhang, H. Jin, D. Meng, J. Wang, and J. Tan, "A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods," *arXiv*, vol. October, 2025, doi: 10.48550/arXiv.2403.02901.

[11] S. Chen, Y. U. Zhang, and Q. Yang, "Multi-Task Learning in Natural Language Processing : An Overview," *ACM Comput. Surv.*, vol. 56, no. 12, 2024, doi: 10.1145/3663363.

[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, doi: 10.1145/3505245.

[13] X. L. Li and P. Liang, "Prefix-Tuning : Optimizing Continuous Prompts for Generation," *arXiv*, vol. Januari, 2021, doi: 10.48550/arXiv.2101.00190.

[14] M. Akter, N. Bansal, and S. K. Karmaker Santu, "Revisiting Automatic Evaluation of Extractive Summarization Task : Can We Do Better than ROUGE ?," *Find. Assoc. Comput. Linguist. ACL 2022*, pp. 1547–1560, 2022, doi: 10.18653/v1/2022.findings-acl.122.

[15] T. Goyal, J. J. Li, and G. Durrett, "News Summarization and Evaluation in the Era of GPT-3," *arXiv*, vol. 23 May, 2022, doi: 10.48550/arXiv.2209.12356.

[16] S. Zhang, D. Wan, and M. Bansal, "Extractive is not Faithful : An Investigation of Broad Unfaithfulness Problems in Extractive Summarization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, vol. 1, pp. 2153–2174. doi: 10.18653/v1/2023.acl-long.120.

[17] Y. Lyu, C. Zhu, T. Xu, Z. Yin, and E. Chen, "Faithful Abstractive Summarization via Fact-aware Consistency-constrained Transformer," *Proc. 31st ACM Int. Conf. Inf. Knowl. Manag.*, pp. 1410–1419, 2022, doi: 10.1145/3511808.3557319.

[18] V. Rennard, G. Shang, J. Hunter, and M. Vazirgiannis, "Abstractive Meeting Summarization : A Survey," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 861–884, 2023, doi: 10.1162/tacl_a_00578.

[19] Y. Mao, X. Ren, H. Ji, and J. Han, "Preserving Factual Consistency with Constrained Generation," *arXiv*, vol. December, 2021, doi: 10.48550/arXiv.2010.12723.

[20] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "applied sciences Abstractive vs . Extractive Summarization : An Experimental Review," *MDPI - Appl. Sci.*, vol. June, 2023, doi: 10.3390/app13137620.

[21] G. Tucudean, M. Bucos, B. Dragulescu, and D. Caleanu, "Natural language processing with transformers : a review," *PeerJ Comput. Sci.*, vol. 7 Agustus, pp. 1–22, 2024, doi: 10.7717/peerj-cs.2222.

[22] W. Phatthiyaphaibun *et al.*, "PyThaiNLP : Thai Natural Language Processing in Python," *arXiv*, vol. December, 2023, doi: 10.48550/arXiv.2312.04649.

[23] Z. Bayramoğlu and M. Uzar, "Performance analysis of rule-based classification and deep learning method for automatic road extraction," *Int. J. Eng. Geosci.*, vol. 8, no. 1, pp. 83–97, 2023, doi: 10.26833/ijeg.1062250.

[24] A. R. Fabbri, W. Kry, B. Mccann, C. Xiong, R. Socher, and D. Radev, "SummEval : Re-evaluating Summarization Evaluation," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 391–409, 2021, doi:

10.1162/tacl_a_00373.
[25] L. Zhang, W. Wang, J. Ma, and Y. Wen, "IWF-TextRank Keyword Extraction Algorithm Modelling," *MDPI - Appl. Sci.*, vol. 14, no. 22, 2024, doi: 10.3390/app142210657.
[26] D. Chen, J. Hu, X. Wei, and E. Wu, "Real3D: The Curious Case of Neural Scene Degeneration," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 1028–1036, 2022.
[27] D. Roy and S. Fakhoury, "Reassessing Automatic Evaluation Metrics for Code Summarization Tasks," in *Pro- ceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, 2021, pp. 1105–1116. doi: 10.1145/3468264.3468588.
[28] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural Abstractive Text Summarization with Sequence-to-Sequence Models," *ACM/IMS Trans. Data Sci.*, vol. 2, no. 1, 2020, doi: 10.1145/3419106.
[29] S. Xu, X. Zhang, Y. Wu, and F. Wei, "Sequence Level Contrastive Learning for Text Summarization," *Proc. AAAI Conf. Artif. Intell.*, 2022, doi: 10.1609/aaai.v36i10.21409.
[30] I. Garrido-muñoz, A. Montejo-ráez, F. Martínez-santiago, and L. A. Ureña-lópez, "applied sciences A Survey on Bias in Deep NLP," *MDPI - Appl. Sci.*, vol. 11, no. 7, pp. 1–26, 2021, doi: 10.3390/app11073184.
[31] T. Sun, J. He, X. Qiu, and X. Huang, "BERTScore is Unfair : On Social Bias in Language Model-Based Metrics for Text Generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3726–3739. doi: 10.18653/v1/2022.emnlp-main.245.
[32] A. Gunakala and A. H. Shahid, "A Comparative Study on Performance of Basic and Ensemble Classifiers WIth Various Datasets," *Appl. Comput. Sci.*, vol. 19, no. 1, pp. 107–132, 2023, doi: 10.35784/acs-2023-08.