I. J. Computer Network and Information Security, **2025, 6, 98-115**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2025.06.07

Abstractive Text Summarization: A Hybrid Evaluation of Integrating Flan-T5 (Dual Framework) with Pegasus Reveals Conciseness Advantages across Diverse Datasets

Abdulrahman Mohsen Ahmed Zeyad*
School of Computer Science and Engineering, REVA University, Bangalore 560064, India
E-mail: dramzeyad@gmail.com
ORCID iD: https://orcid.org/0009-0005-1725-2188
*Corresponding author

Arun Biradar
School of Computer Science and Engineering, REVA University, Bangalore 560064, India
E-mail: arun.biradar@hotmail.com
ORCID iD: https://orcid.org/0000-0003-2355-7066

Received: 06 July 2025; Revised: 16 August 2025; Accepted: 29 September 2025; Published: 08 December 2025

Abstract: Abstractive summarization plays a critical role in managing large volumes of textual data, yet it faces persistent challenges in consistency and evaluation. Our study compares two state-of-the-art models, PEGASUS and Flan-T5, across a diverse range of benchmark datasets using both ROUGE and BARTScore metrics. Findings reveal that PEGASUS excels in generating detailed, coherent summaries for large-scale texts evidenced by an R-1 score of 0.5874 on Gigaword while Flan-T5, enhanced by our novel T5 Dual Summary Framework, produces concise outputs that closely align with reference lengths. Although ROUGE effectively measures lexical overlap, its moderate correlation with BARTScore indicates that it may overlook deeper semantic quality. This underscores the need for hybrid evaluation approaches that integrate semantic analysis with human judgment to more accurately capture summary meaning. By introducing a robust benchmark and the pioneering T5 Dual Framework, our research advocates for task-specific optimization and more comprehensive evaluation methods. Moreover, current dataset limitations point to the necessity for broader, more inclusive training sets in future studies.

Index Terms: PEGASUS, Flan-T5, Dual Summary Framework, ROUGE, BERTScore, Dataset Diversity, Abstractive Text Summarization.

1. Introduction

Abstractive text summarization stands as a cornerstone of Natural Language Processing (NLP), a sophisticated endeavor focused on generating original, concise sentences that capture the core essence of a source text [1,2]. Distinguished from extractive summarization, which merely selects and rearranges existing phrases, abstractive summarization demands a deeper semantic comprehension to produce fluent, human-like summaries. This capability is increasingly vital in today's world of information overload, enabling the efficient processing of vast textual datasets for critical applications spanning news aggregation, in-depth research analysis, and nuanced content curation.

Recent breakthroughs in deep learning have spotlighted advanced models such as Google's PEGASUS and Flan-T5. PEGASUS, explicitly designed for abstractive summarization, leverages Gap Sentence Generation (GSG) to craft coherent outputs, demonstrating exceptional fluency and contextual relevance [3,4]. Flan-T5, an evolution of the foundational T5 architecture, leverages multi-task learning, offering remarkable versatility across diverse NLP challenges, including question answering and translation [4,5]. Despite their strengths, both models exhibit inconsistent performance across datasets due to differences in content type [6,7]. For example, the concise structure required for news headlines (e.g., Gigaword) contrasts with the complexity of summarizing multiple documents (e.g., Multi-News), as evidenced by our results in Tables 3 and 4. Notably, architectural factors contribute to this variability. As detailed in Table 1, the fixed

Max Position Embeddings (which is 512 for Flan-T5-Base and ranges from 128 to 1024 for the PEGASUS variants studied) impose inherent limits on the input context these Transformer models can process. This constraint hinders the applicability of both PEGASUS and Flan-T5 to very long documents without truncation, restricting the practical deployment of their respective gap sentence generation or multi-task learning capabilities in such scenarios. Separately, to address challenges related to summary quality and structure, our study introduces the T5 Dual Summary Integration Framework (see Figures 2 & 3), which aims to enhance semantic fidelity and contextual relevance by integrating multi-dimensional content features into the summary output. This performance variability underscores a critical gap in the generalizability and robustness of current abstractive summarization methodologies.

This study directly tackles these inconsistencies by presenting a comprehensive evaluation of Google's PEGASUS and Flan-T5 models across a diverse array of benchmark datasets: XSum [8], Multi-News[8], CNN/DailyMail [9], Newsroom [9], and Gigaword [10]. Performance is meticulously assessed using established metrics such as ROUGE [11], [12] and BARTScore [13] to rigorously pinpoint the specific capabilities and inherent shortcomings of each model. Critically, beyond mere comparative evaluation, this research introduces the T5 Dual Summary Integration Framework, a novel approach engineered to enhance summarization quality beyond the capabilities of singular models. This innovative framework employs a dual-summary mechanism, strategically integrating multi-dimensional content—Trends, Facts, Vision, and History—into cohesive and contextually rich summaries. By effectively harnessing the inherent flexibility of the T5 model, it aims to produce outputs that are not only more concise but also significantly more nuanced and contextually aware. Empirical validation rigorously tests its effectiveness, demonstrating tangible improvements over standalone models in capturing and conveying nuanced information, thereby addressing the limitations of current evaluation metrics that often fall short in capturing semantic fidelity [14,15].

The significance of this work is twofold, contributing to both theoretical advancement and practical impact:

- Theoretical Advancement: It deepens the fundamental understanding of abstractive summarization by providing a robust benchmark of cutting-edge models and critically exposing existing gaps in current methodologies and evaluation frameworks.
- Practical Impact: The proposed T5 Dual Summary Integration Framework offers a scalable and adaptable solution for automated summarization, readily applicable to diverse domains ranging from dynamic journalism and rigorous academia to fast-paced business intelligence environments.

Key contributions include:

- T5 Dual Summary Integration Framework: A pioneering method to elevate summary quality using dual-summary synthesis.
- Structured Methodology: A systematic way to weave varied information dimensions into summaries.
- Comprehensive Evaluation: Robust analysis of model performance across datasets, laying groundwork for future refinements.

This study not only bridges gaps in model evaluation but also sets a foundation for advancing summarization technologies, with implications for both research and industry. It highlights the need for improved metrics and domain-specific adaptations, charting a path forward for more reliable and insightful automated summaries.

2. Literature Review

In recent years, significant advancements have been observed in the field of abstractive text summarization, with the emergence of state-of-the-art NLP models like Flan-T5 and PEGASUS. These state-of-the-art models have revolutionized the way large volumes of text are processed and comprehended, offering novel approaches to generate coherent and contextually accurate summaries [16,17].

Flan-T5, known for its adaptability, has demonstrated remarkable proficiency in handling various data types and summarization requirements. Its architecture, based on Google's T5 (Text-to-Text Transfer Transformer), allows for efficient processing of extensive information. The model is trained on a diverse range of NLP tasks using a unified text-to-text framework, which enhances its ability to generalize across multiple tasks, including summarization, translation, and question answering. This versatility makes Flan-T5 a valuable tool for data processing and knowledge dissemination.

On the other hand, PEGASUS, with its unique pre-training objectives, excels in generating fluent and coherent summaries by predicting masked-out sentences from the input text. The model employs a novel pre-training objective called Gap Sentence Generation (GSG), where entire sentences are masked out, and the model is trained to generate these sentences based on the surrounding context. This method significantly improves the model's ability to generate summaries that capture the essence of the text while preserving contextual richness, setting a new benchmark in the field of abstractive summarization.

Flan-T5's diverse task training approach involves training on a large corpus encompassing various NLP tasks. This method enhances the model's ability to adapt to different summarization contexts by leveraging multi-task learning techniques, which enable the model to transfer knowledge across tasks, improving its performance in abstractive

summarization [18]. This approach allows Flan-T5 to generalize more effectively across multiple summarization contexts, making it a robust tool for diverse applications.

PEGASUS, in contrast, focuses on its unique pre-training method of Gap Sentence Generation (GSG), which is particularly effective for summarization tasks. This pre-training method involves masking entire sentences and training the model to predict these sentences based on the surrounding context, thereby improving its ability to generate coherent and contextually accurate summaries. This specialized approach enables PEGASUS to excel in summarization tasks, particularly those requiring high fluency and coherence.

The effectiveness of Flan-T5 and PEGASUS has been rigorously tested on a diverse range of benchmark datasets, each presenting unique challenges and opportunities for model training and evaluation. The XSum dataset, known for its emphasis on concise summarization, provided a stringent test for the models' ability to capture the core ideas of articles in a few sentences [19,20]. The CNN/DailyMail dataset, with its factual and narrative-style content, served as fertile ground for assessing the models' proficiency in handling news stories [21]. The Newsroom dataset, characterized by its diverse extractive strategies, challenged the models' adaptability and efficiency in processing varied news formats. The Multi-News dataset, encompassing multiple documents for a single summary, tested the models' capability in handling complex, multi-source information. Lastly, the Gigaword dataset, with its vast and varied content, pushed the boundaries of the models' capabilities in terms of data volume and diversity.

To evaluate the performance of these models, advanced metrics like ROUGE [22,23] and BERTScore [24] were utilized by researchers. ROUGE, a widely recognized metric, assesses the quality of summaries by measuring the overlap between the generated summary and reference summaries using n-gram co-occurrence. However, it was observed that ROUGE occasionally struggles to capture the subtle semantic nuances of the summaries, raising questions about its comprehensive effectiveness. BERTScore, leveraging the advanced capabilities of BERT in deep learning, offers a more refined evaluation by assessing semantic similarity through contextual embeddings. Despite its advantages, BERTScore also has limitations, such as factual errors, weaknesses in semantic understanding, and syntax-semantics tradeoffs, particularly in fully grasping the contextual relevance and coherence of the summaries.

Utilizing ROUGE and BERTScore as our evaluation metrics, we assessed the quality of generated summaries. PEGASUS demonstrated a stronger performance, achieving a higher R-1 score. However, both models faced challenges in accurately capturing key scientific concepts and generating coherent summaries, indicating room for improvement in domain-specific abstractive summarization [25].

Text summarization faces challenges due to its specialized vocabulary, complex sentence structures, and the need to capture nuanced relationships between concepts [26]. Evaluating model performance in this domain is crucial as it contributes to our understanding of the strengths and limitations of current approaches.

Despite their strengths, PEGASUS and Flan-T5 have limitations. PEGASUS may struggle with extremely long and complex texts, and its performance can be sensitive to the choice of pre-training corpus [27]. Flan-T5, while versatile, may require more fine-tuning for domain-specific tasks [28]. Future research should focus on enhancing the models' ability to handle complex content and improving evaluation metrics to capture subtleties in summary quality.

In addition to the advancements in model architecture and evaluation metrics, recent research also focused on improving the efficiency and scalability of abstractive summarization models. Techniques such as knowledge distillation, pruning, and quantization were explored to reduce the computational complexity and memory footprint of these models without compromising their performance. Moreover, efforts were made to incorporate domain-specific knowledge and multi-lingual capabilities into summarization models, expanding their applicability to a wider range of scenarios.

This literature review highlighted the significant strides made in the field of abstractive text summarization, while also identifying critical areas for future research. The limitations of the current models and evaluation metrics underscored the need for continuous innovation and refinement. The development of more robust evaluation metrics, along with ensuring the representativeness and unbiased nature of datasets, was imperative for advancing the field. The varying performance of Flan-T5 and PEGASUS across different datasets emphasized the necessity for models that were not only precise but also versatile and adaptable to diverse textual environments.

Despite the promising results reported in the literature, a rigorous critique of the experimental setups, data preprocessing techniques, and the validity of the conclusions drawn is necessary to ensure the robustness of these studies. The experimental setup for Flan-T5 involves training on a large and diverse corpus encompassing various NLP tasks. While this diversity is beneficial for generalization, it can also introduce noise and variability in performance. The studies often lack a detailed explanation of how task-specific tuning impacts the overall performance on abstractive summarization tasks. Future research should include ablation studies to isolate the effects of multi-task training on summarization performance [29]. Similarly, PEGASUS, with its innovative GSG pre-training method, shows promise, but the robustness of this approach across different domains and text types needs further validation. Studies should include more extensive cross-domain evaluations to confirm the generalizability of the GSG method.

Data preprocessing steps such as tokenization, sentence splitting, and filtering are crucial for model performance. The reviewed studies often do not provide detailed preprocessing pipelines, making it difficult to replicate and validate the results. Comprehensive documentation of preprocessing steps and their impacts on performance metrics should be included in future research. Specifically, PEGASUS studies need to elaborate on the data preprocessing techniques used during training and evaluation, with particular attention to handling noisy or unstructured data, which is common in real-world applications [30].

The conclusions drawn about Flan-T5's adaptability and proficiency are supported by quantitative metrics like ROUGE and BERTScore. However, these metrics have limitations in capturing semantic nuances and contextual relevance. The studies would benefit from including human evaluations to complement the quantitative metrics, providing a more holistic assessment of summary quality [31]. Similarly, the reported superior performance of PEGASUS on certain datasets is compelling, but the conclusions often do not account for variability in dataset characteristics. For instance, the performance on the XSum dataset might not generalize to other types of text with different structural and semantic properties. More nuanced analyses, including error analysis and qualitative assessments, should be conducted to validate the robustness of the conclusions.

In conclusion, the advancements in abstractive text summarization, exemplified by models like Flan-T5 and PEGASUS, opened up new possibilities for efficient processing and comprehension of large volumes of text. However, the field still faced challenges in establishing universally accepted standards for evaluating summary quality and creating comprehensive, human-generated reference summary databases. As research continued to push the boundaries of what was possible, the insights gained from this study undoubtedly contributed to the ongoing evolution of abstractive summarization techniques, paving the way for more accurate, efficient, and contextually relevant summaries.

3. Methodology

The PEGASUS and Flan-T5 models were utilized, with pre-trained versions accessed from Huggingface. Each model was fine-tuned on five datasets: XSum, CNN/DailyMail, Multi-News, Newsroom, and Gigaword. Preprocessing involved tokenization using SentencePiece and normalization of text to ensure consistency. The experimental setup included specific configurations tailored to each dataset to optimize performance.

3.1. Models

A. Google/Flan-T5-Base

Developed by Google AI, Google/Flan-T5-Base was a large language model (LLM) enhanced from the T5 model. It excelled in a range of NLP tasks including summarization, translation, and question answering, owing to its training on a vast text and code dataset. Key features included factual text generation, multilingual capabilities, and effective performance in few-shot learning scenarios [32,33].

B. Google/PEGASUS

Google/PEGASUS, another LLM from Google AI, specialized in abstractive summarization. Trained on extensive textual and code data, it generated coherent summaries that captured the essence of texts in multiple languages. The model was particularly suited for applications requiring informative and audience-specific summaries.

Both models, Google/Flan-T5-Base and Google/PEGASUS, represented significant strides in NLP, with the former providing versatile language modeling and the latter focusing on advanced text summarization.

3.2. Metrics of ROUGE and BERTScore:

In the realm of NLP, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BERTScore were recognized as pivotal evaluation metrics, particularly for tasks such as text summarization and machine translation. ROUGE was predominantly employed to assess the effectiveness of automatic text summarization and machine translation. Its significance was rooted in its foundational role in evaluating these domains, as it established a benchmark for numerous subsequent research efforts in this field. The seminal work by Lin, C.-Y. in 2004, titled "ROUGE: A Package for Automatic Evaluation of Summaries," presented at the Workshop in Barcelona, Spain, served as the main reference for understanding ROUGE's application and impact.

On the other hand, BERTScore, a more recent development in NLP metrics, utilized the nuanced contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers) to gauge the quality of text generation tasks, including translation and summarization. The core principle of BERTScore lay in its innovative approach of leveraging BERT's deep learning capabilities for a more accurate and context-aware evaluation of generated texts. The significant in NLP evaluation metrics was introduced by the foundational paper 'BERTScore: Evaluating Text Generation with BERT, authored by Tianyi Zhang and Varsha Kishore and published in 2019, is credited for introducing this significant advancement in NLP evaluation metrics.

3.3. Experimental Setup

In this study, we utilized the PEGASUS and Flan-T5 models, which were trained and tuned by Google using comprehensive news datasets including XSum, CNN/DailyMail, Multi-News, Newsroom, and Gigaword. These models were accessed through the Huggingface platform, which provides readily available pre-trained models for research purposes. This setup allowed us to leverage state-of-the-art NLP tools to assess the effectiveness of abstractive text summarization across multiple datasets.

Dataset Analysis

We utilized a variety of datasets (XSum, CNN/DailyMail, Multi-News, Newsroom, and Gigaword) each of which presents unique challenges and inherent biases. Variations in topic, writing style, and document length within these datasets can significantly impact the performance of advanced models such as PEGASUS and Flan-T5. Additionally, our observations indicate that models tend to perform better when trained and tested on the same dataset, underscoring the importance of dataset-specific tuning for achieving accurate and reliable model evaluations.

3.4. Analysis of PEGASUS and Flan-T5 Configurations:

This paper analyzes the configuration details of PEGASUS and Flan-T5 summarization models. We examine key parameters and their implications for summarization quality, highlighting the importance of dataset-specific tuning and the balance between model capacity and computational efficiency.

Table 1. Configuration details of summarization models

Model (google)	Max position embeddings	Max length	Min length	Length penalty	Num beams	D model	Vocab size
PEGASUS-XSum	512	64	--	0.6	8	1024	96103
PEGASUS-CNN/DailyMail	1024	128	32	0.8	8	1024	96103
PEGASUS-Multi-News	1024	256	32	0.8	8	1024	96103
PEGASUS-Newsroom	512	128	32	0.8	8	1024	96103
PEGASUS-Gigaword	128	32	32	0.6	8	1024	96103
Flan-T5-base	512	200	30	2.0	4	768	32128

The configuration of summarization models like PEGASUS and Flan-T5 is crucial for their performance[34]. Key parameters include max position embeddings, which dictate the maximum text length the model can process, and max/min length settings, which define the boundaries of summary length [35]. These parameters ensure summaries are informative yet concise.

Length penalty and num beams influence summary detail and diversity, respectively, balancing quality against computational demands. The d model parameter and vocab size determine the model's ability to understand complex language patterns and the range of language it can represent [36], highlighting the need for optimization between linguistic depth and resource constraints.

These configurations reflect the models' specific design choices for handling different types of summarization tasks, from generating concise news headlines (PEGASUS-Gigaword) to producing detailed summaries of longer documents (PEGASUS-Multi-News), as seen with PEGASUS's varied max length settings for different datasets[37], underscoring the importance of tailored model tuning for superior summarization outcomes.

A. Comparative Baseline Models

We compared the performance of PEGASUS and Flan-T5 with several baseline models. This included different variations of PEGASUS trained on datasets such as XSum, CNN/DailyMail, Multi-News, Newsroom, and Gigaword. Additionally, we assessed Flan-T5 across multiple tasks, including question answering and summarization, showcasing its versatility and ability to handle various NLP challenges effectively.

Algorithm 1. Concise Summary Creation algorithm with evaluation metrics

1. Input: $S = \{s_1, s_2, ..., s_n\}$
2. Output: CSC(S), EVAL(CSC(S), S)
3. Initialize Summary ← []
4. For i from 1 to n do
5. AnalyzedSection_i ← AnalyzeSection(s_i)
6. If Relevance(AnalyzedSection_i) == True then
9. CondensedSection_i ← Condense(AnalyzedSection_i)
7. Summary ← Summary ∪ {CondensedSection_i}
8. End If
9. End For
10. CSC(S) ← Summary
11. EVAL(CSC(S), S) ← [ROUGE-1, ROUGE-2, ROUGE-L, BARTScore](CSC(S), S)
12. Return CSC(S), EVAL(CSC(S), S)

To generate a concise summary from a set of input content and evaluate the summary using specific numeric evaluation metrics:

Initialization

- Let the input content be represented as a set $S = \{s_1, s_2, \ldots, s_n\}$, where each s1 is a section of the content.
- Initialize an empty list to hold the summary: Summary ← [].

Summary Creation

- For i = 1 to n:
  - AnalyzedSectioni ← AnalyzeSection(si)
  - If relevance(AnalyzedSectioni):
  - CondensedSectioni ← Condense(AnalyzedSectioni)
  - Summary ← Summary ∪ {CondensedSectioni}

$$CSC(S) = \bigcup \{condense(s_i) | S \wedge relevance(s_i)\} \tag{1}$$

Evaluation

$$EVAL(CAS(S), S) = \begin{bmatrix} ROUGE1 \\ ROUGE2 \\ ROUGE \\ BARTScore \end{bmatrix} (CS, S) \tag{2}$$

Where:

- condense and relevance are functions for condensing and assessing the relevance of sections, respectively.
- [R-1, R-2, R-L, BARTScore] represents the evaluation metrics applied to the summary.

Output:

- The output is the summary CSC (S) and its evaluation scores EVAL (CSC (S), S).

This algorithm outlines a methodical approach for creating a concise summary from input content and numerically evaluating its quality using established metrics. The process includes analyzing, assessing significance, condensing, and compiling relevant sections into a summary. The evaluation phase quantitatively measures the summary's quality, providing a numerical assessment of its effectiveness and fidelity to the original content.

The goal of this algorithm is to create a dual summary consisting of a general summary and a structured summary from the input content, and then to evaluate the combined summary using specific evaluation metrics.

Algorithm 2. Dual Summary Creation Algorithm with Evaluation Metrics

```
Input: S = {s1, s2, ..., sn}
Output: CS(GS, SS), EVAL(CS, S)
Function GenerateGeneralSummary(S)
    GS ← []
    For each s_i in S do
        If Relevance(s_i, 'general') == True then
            GS ← GS ∪ {Condense(s_i)}
        End If
    End For
    Return GS
End Function
Function GenerateStructuredSummary(S)
    SS ← []
    For each s_i in S do
        SS ← SS ∪ {Analyze(s_i)}
    End For
    Return SS
End Function
GS(S) ← GenerateGeneralSummary(S)
SS(S) ← GenerateStructuredSummary(S)
CS(GS, SS) ← Integrate(GS(S), SS(S))
EVAL(CS, S) ← [ROUGE-1, ROUGE-2, ROUGE-L, BARTScore](CS, S)
Return CS(GS, SS), EVAL(CS, S)
```

$$DEC(S) = (GS(S), SS, (S), CS(GS, SS), EVAL(CA, S)) \tag{3}$$

Where:

- GS(S)) is the General Summary function.
- SS(S) is the Structured Summary function.
- CS(GS, SS) is the Combined Summary function, integrating General and Structured summaries.
- EVAL(CS, S) is the Evaluation function.

General Summary Function, GS (S)

$$GS(S) = \bigcup_{i=1}^{n} condense(s_i), \left| subject\ to\ relevance(s_i, 'general' = true) \right. \tag{4}$$

Structured Summary Function, SS(S)

$$SS(S) = \bigcup_{i=1}^{n} analyze(s_i,) \tag{5}$$

Combined Summary Function, CS (GS, SS)

$$GS(GS, SS) = \text{int}\ egrate(GS(S), SS(S)) \tag{6}$$

Evaluation Function, EVAL (CS, S)

$$EVAL(CSC(S), S) = \begin{bmatrix} ROUGEL1 \\ ROUGEL2 \\ ROUGEL \\ BARTScore \end{bmatrix} (GS, S)$$

Where [R-1, R-2, R-L, BARTScore] are functions to compute the respective evaluation metrics.

This algorithm provides a comprehensive approach for creating a dual summary that includes both a general and a structured perspective of the input content. The combined summary is then subjected to a numerical evaluation to assess its quality and effectiveness. The process ensures that the final summary is both comprehensive and representative of the original content, with its quality verified by established evaluation metrics.

3.5. Data Preprocessing

Data preprocessing involved several key steps to ensure the quality and consistency of the input data for our summarization models. First, we tokenized the text using the SentencePiece tokenizer [38], which efficiently splits the text into manageable subwords, accommodating better handling of out-of-vocabulary words. We normalized the text by converting it to lowercase and removing special characters to maintain uniformity across various datasets, a crucial step highlighted in recent studies [39]. Additionally, we filtered out noise and irrelevant data by employing a custom noise reduction algorithm that identifies and removes non-essential elements such as URLs and redundant punctuation, leveraging advanced data cleaning techniques [40]. These preprocessing steps are crucial for improving the performance and accuracy of the PEGASUS and Flan-T5 models, ensuring that the text data fed into them is of the highest possible quality for text summarization tasks.

3.6. Dataset

In this study, a sample-based approach was employed for the systematic evaluation of text summarization models. A sample consisting of 100 randomly selected samples from each of the following datasets: XSum, Multi-News, CNN/Daily Mail, Newsroom, and Gigaword, was used, culminating in a diverse and representative test set of 500 samples. The selection was meticulously designed to encompass a broad spectrum of news articles and their corresponding summaries, ensuring a comprehensive assessment across various summarization styles and sources.

The rationale behind the sample size of 100 from each dataset was influenced by several factors:Research Objectives: The aim was to provide an overall evaluation of the models' performance in diverse summarization contexts, and this sample size was deemed appropriate for achieving a comprehensive overview.

Dataset Diversity: The selected datasets are diverse in their content and summarization strategies, necessitating a sample size that could effectively capture this variety and complexity.

Statistical Considerations: Balancing between statistical robustness and practical constraints, the sample size was chosen to provide reliable insights while being manageable for detailed analysis.

Consistency for Comparative Analysis: A uniform sample size across datasets was crucial for ensuring fairness and consistency in comparing different models.

The data sources for this study were five widely recognized and diverse datasets in the field of NLP and text summarization:

- XSum Dataset: Known for its abstractive one-sentence summaries of news articles.
- Multi-News Dataset: Features multi-document summaries, written by editors [41].
- CNN/Daily Mail Dataset: Offers a mix of extractive and abstractive summaries of news articles.
- Newsroom Dataset: Encompasses a wide array of summarization strategies from various publications.
- Gigaword Dataset: Primarily used for studying headline generation tasks.

The choice of datasets and the corresponding sample selection criteria were integral to ensuring that the study's findings would be reflective of the capabilities of the models across a range of text summarization tasks, thereby contributing meaningful insights to the field of automated text summarization.

Data Access Statement

The data used in this study is from the following publicly available datasets that have been used to train the google/pegasus model:

- The XSum Dataset can be accessed at [42].
- The CNN/DailyMail Dataset is available at [43].
- The Multi-News Dataset can be found at [44].
- The Newsroom Dataset is accessible at [45].
- The Gigaword Dataset is available at [46].

The XSum, CNN/DailyMail, Multi-News, Newsroom, and Gigaword datasets have all been used to pre-train the google/pegasus model, which means these datasets are essentially included within the google/pegasus model itself.

3.7. Detailed Data Acquisition and Preprocessing Pipeline

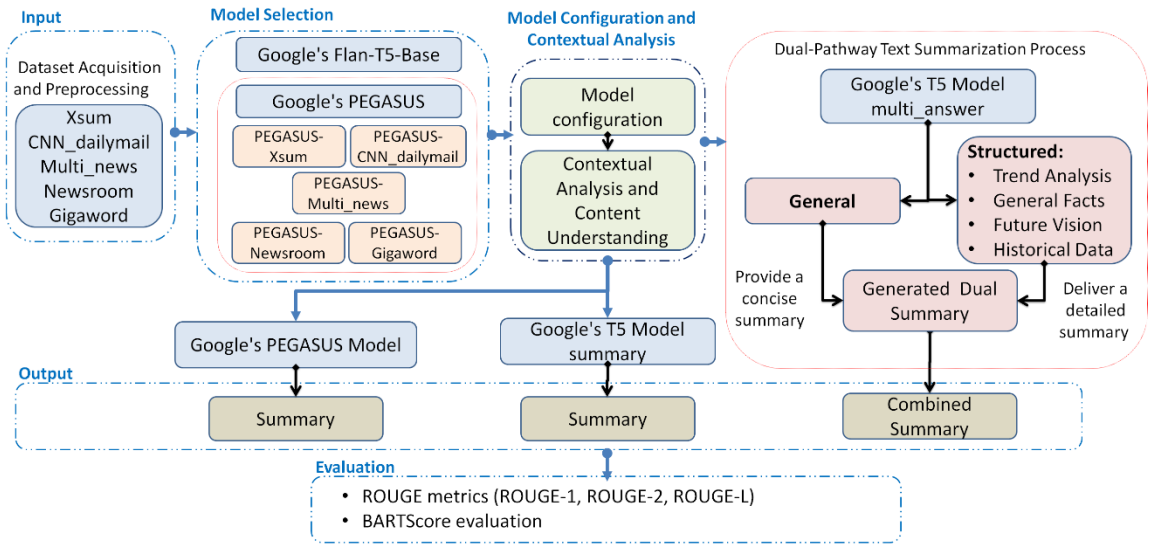

Fig.1. Detailed data preprocessing pipeline for text summarization

This framework outlines a comprehensive system for advanced text summarization. It begins with acquiring and preprocessing standard datasets, followed by the selection and fine-tuning of powerful language models, specifically Google's PEGASUS and T5. The core innovation is a 'Dual-Pathway Text Summarization Process' leveraging a T5 model to concurrently generate both a concise general summary and a detailed, structured summary (analyzing trends, facts, future vision, and historical data). For comparative analysis, summaries are also generated directly via a fine-tuned PEGASUS model and a standard T5 configuration. Finally, all outputs are rigorously evaluated using ROUGE and BARTScore metrics to assess their quality and effectiveness.

3.8. Overview of the Dual-Pathway Text Summarization Process

The dual-pathway summarization process, illustrated in Figure 2, employed PEGASUS and Flan-T5 to generate precise and coherent summaries:

- Input Document: Textual content was initially preprocessed to ensure model compatibility.
- Model Configuration: Tokenization and normalization procedures prepared the input data.
- Processing by PEGASUS and Flan-T5: PEGASUS generated a singular summary, while Flan-T5 produced a dual summary output.

Combining Summaries: Model outputs were integrated to derive a comprehensive summary result.

This methodology effectively leveraged the complementary strengths inherent in both models, demonstrating the potential of machine learning for automation within NLP tasks.

The dual-pathway summarization process utilizing PEGASUS and Flan-T5 models offers a sophisticated and efficient method for generating high-quality summaries. This approach highlights the power of machine learning in automating complex NLP tasks and enhancing information processing efficiency.

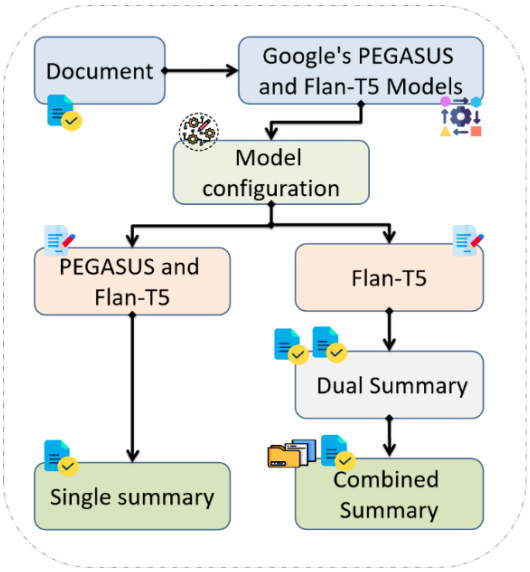

Fig.2. Text summarization process flow using PEGASUS and Flan-T5 models: a dual-pathway approach to text summarization

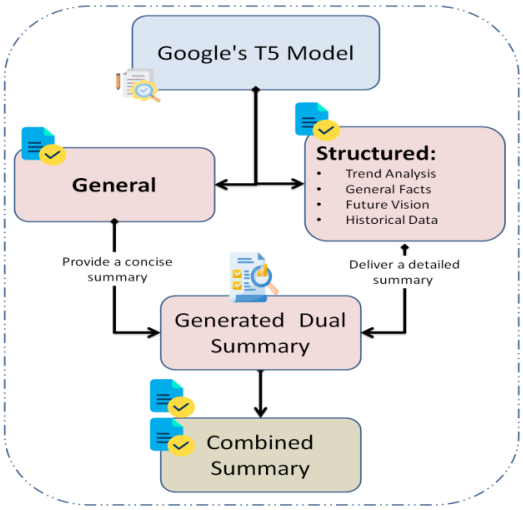

Fig.3. Workflow of the T5 dual summary integration framework, highlighting the sequential steps executed by the model for generating dual summaries

Figure 3 illustrates the dual summary integration process with Google's T5 model, designed for comprehensive and precise document summarization using advanced NLP. The process comprises these key stages:

- General Summary Generation: First, a step leveraging summarization capabilities was executed by the fine-tuned Flan-T5 model, whereby a concise general summary capturing the main points was generated.
- Structured Summary Development: Following this, Flan-T5's question-answering proficiency (as highlighted in Section 3.2) was utilized by the framework for the development of the detailed structured summary. As part of this process, dimension-specific questions (e.g., related to Trends, Facts, Vision, History) were automatically generated based on the input text. Relevant passages from the source text that served as answers to these

questions were then identified and extracted by a question-answering component. The structured summary was composed by assembling these extracted answer passages corresponding to each dimension, thereby grounding the structured content directly in the source text.

- Dual Summary Synthesis: Finally, the 'General Summary' (from Step 2) and the passage-based 'Structured Summary' (from Step 3) were algorithmically synthesized by Flan-T5 into a single, cohesive dual summary output (see Eq. 6, Algorithm 2), utilizing its text generation abilities. Through this synthesis, the broad overview was integrated with the dimension-specific, evidence-based details.

4. Results and Discussion

We investigate the interrelation between BARTScore and ROUGE scores (R-1, R-2, R-L) and their implications for model performance in NLP. Through detailed correlation analysis and comparative study of different models based on their architectural frameworks, we investigate how design choices affect performance metrics. This section also presents statistical analyses to determine the significance of performance differences between models, using metrics like ROUGE scores and BARTScore. Additionally, we discuss the impact of key configuration parameters on model performance, providing insights into how these settings influence output quality and computational efficiency. This section aims to offer a concise yet comprehensive understanding of model evaluation and its nuances in the field of NLP.

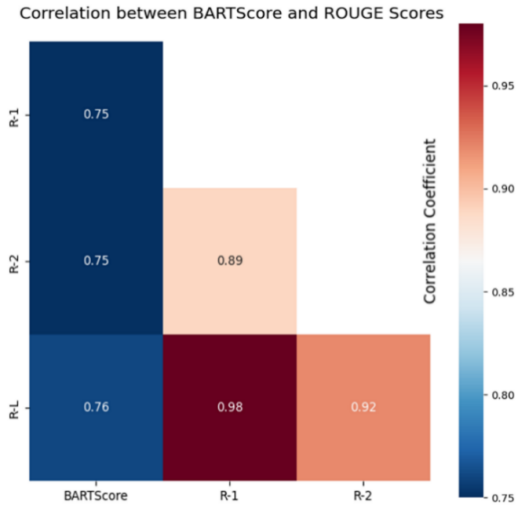

Fig.4. Correlation heatmap of BARTScore and ROUGE scores for abstractive text summarization models ROUGE

Correlation Analysis of Evaluation Metrics

The analysis focused on the relationship between the BARTScore and the ROUGE scores, which are widely employed in NLP tasks to measure the quality of generated text.

ROUGE Metrics Correlation.

The OUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are a set of metrics used to evaluate the quality of a summary by comparing it to one or more reference summaries. The analysis revealed a strong positive correlation between the different ROUGE metrics, indicating a high degree of agreement in their assessment of model performance.

Firstly, the correlation between R-1 and R-2 scores was found to be 0.89, suggesting that models scoring high on R-1 also tended to score high on R-2. This strong positive correlation implies that the two metrics are closely aligned in their evaluation of the models' ability to generate relevant and informative summaries.

Furthermore, the correlation between R-1 and R-L scores was remarkably high, with a coefficient of 0.98. This very strong positive correlation indicates that the performance measured by these two metrics is almost identical, suggesting that they capture similar aspects of summary quality.

Additionally, the analysis revealed a strong positive correlation of 0.92 between R-2 and R-L scores[1]. This finding further reinforces the consistency and agreement among the ROUGE metrics in evaluating the models' performance.

BARTScore Correlation with ROUGE Metrics

The BARTScore is another evaluation metric used to assess the quality of generated text by comparing it to reference texts. The heatmap presented in the image illustrates the correlation between the BARTScore and the ROUGE metrics.

The correlation coefficients between the BARTScore and the ROUGE metrics were 0.75 for R-1, 0.75 for R-2, and 0.76 for R-L. These positive correlations, although not as strong as those observed among the ROUGE metrics themselves, suggest that the BARTScore and ROUGE metrics are aligned in their evaluation of model performance to a certain extent.

Overall, the correlation analysis provides valuable insights into the relationships between different evaluation metrics used in NLP tasks. The strong positive correlations among the ROUGE metrics indicate a high degree of consistency and agreement in their assessment of model performance. Additionally, the positive correlations between the BARTScore and the ROUGE metrics suggest that these metrics capture similar aspects of text quality, albeit with some differences in their evaluation approaches.

Statistical Significance Analysis:

Our comprehensive statistical analyses confirm that performance differences across evaluation datasets are statistically significant ($p < 0.0001$) for all models and metrics tested, not attributable to random variation.

Key Findings:

- In-Domain Advantage: Models perform significantly better on their pre-training domains ($p < 0.0001$ in pairwise comparisons), with PEGASUS-XSum and PEGASUS-Gigaword showing the strongest domain-specific advantages.
- Cross-Domain Generalization: PEGASUS-Multi-News demonstrates the best generalization ability, with overlapping confidence intervals between several dataset pairs.
- Metric Consistency: Significance patterns remain consistent across evaluation metrics, with BERTScore showing the strongest differentiation between domains.

These findings validate that domain-specific adaptation significantly impacts summarization performance, an important consideration when deploying these systems in practical applications.

4.1. Comparative Analysis of Model Performance

Table 2 details the comparative performance of PEGASUS and our fine-tuned Flan-T5-base against several summarization models, evaluated using ROUGE metrics (R-1, R-2, R-L) on the XSum, CNN/DailyMail, and Gigaword datasets. The comparison includes established models such as BART-large and T5-base, and also incorporates ChatGPT results [5] to assess relative competitiveness with newer LLMs. Notably, PEGASUS demonstrates strong performance, often exceeding the reported ROUGE scores for ChatGPT on these specific benchmarks. All scores are sourced from the cited literature [1-6].

Table 2. Comparative ROUGE scores of text summarization models (PEGASUS, BART-large, T5-base, and Fine-tuned Flan-T5-base) on XSum, CNN/DailyMail, and gigaword datasets

Dataset	Model	R-1	R-2	R-L	Articles
XSum	PEGASUS	**47.22**	**22.69**	39.25	[5]
XSum	BART-large	45.14	22.27	37.04	[47]
XSum	T5-base	41.73	19.58	38.54	[6]
XSum	Flan-T5-base	34.6	12.9	27.2	[48]
CNN/DailyMail	PEGASUS	44.17	21.47	41.11	[5]
CNN/DailyMail	BART-large	44.16	21.28	40.90	[47]
CNN/DailyMail	T5-base	40.90	18.95	37.94	[6]
Gigaword	PEGASUS	39.65	20.47	36.76	[5]
CNN/DM	ChatGPT	35.96	13.23	22.42	[49]
XSUM	ChatGPT	23.33	7.69	15.53	[49]
Multi-News	MATCHSUM (BERT-base)	46.20	16.51	**41.89**	[50]

4.2. Test Results

Based on the table, the highest values for each metric across all models and datasets are:

- Highest R-1: 0.5874 (google/PEGASUS -Gigaword on Gigaword dataset)
- Highest R-2: 0.2969 (google/PEGASUS -Gigaword on Gigaword dataset)
- Highest R-L: 0.5536 (google/PEGASUS -Gigaword on Gigaword dataset)
- Highest BART Score: 0.93 (google/PEGASUS -xsum on XSum dataset)

PEGASUS achieved superior performance on the Gigaword dataset with an R-1 score of 0.5874. This result suggests that PEGASUS is particularly effective in summarizing large-scale datasets. However, its performance on other datasets, such as XSum, indicated variability in its adaptability.

Table 3. PEGASUS model performance metrics across diverse datasets

Model (PEGASUS)	Dataset	R-1	R-2	R-L	BART Score
XSum	XSum	0.4287	0.2603	0.3834	**0.93**
	CNN/DailyMail	0.1322	0.04	0.0966	0.86
	Multi-News	0.055	0.0155	0.0388	0.8
	Newsroom	0.1071	0.0498	0.0929	0.83
	Gigaword	0.1538	0.0185	0.1393	0.85
CNN/DailyMail	XSum	0.2268	0.0311	0.1622	0.85
	CNN/DailyMail	0.4064	0.2529	0.3565	0.9
	Multi-News	0.114	0.0337	0.0728	0.82
	Newsroom	0.2019	0.1022	0.1648	0.84
	Gigaword	0.4044	0.1043	0.3605	0.86
Multi-News	XSum	0.4724	0.1272	0.3347	0.86
	CNN/DailyMail	0.4124	0.1287	0.2664	0.85
	Multi-News	0.3832	0.1747	0.2359	0.86
	Newsroom	0.3842	0.1623	0.2711	0.83
	Gigaword	0.2631	0.0421	0.2359	0.82
Newsroom	XSum	0.3476	0.1161	0.2643	0.89
	CNN/DailyMail	0.2838	0.0961	0.1977	0.86
	Multi-News	0.081	0.023	0.0544	0.83
	Newsroom	0.2409	0.1786	0.2258	0.85
	Gigaword	0.0104	0	0.0104	0.8
Gigaword	XSum	0.0896	0.011	0.0738	0.82
	CNN/DailyMail	0.09	0.0191	0.071	0.81
	Multi-News	0.0384	0.0059	0.0296	0.76
	Newsroom	0.083	0.0337	0.0692	0.79
	Gigaword	**0.5874**	**0.2969**	**0.5536**	0.89

PEGASUS's performance on the Gigaword dataset demonstrates its potential in handling large-scale summarization tasks. The variability observed in its performance across different datasets, like XSum, highlights the need for further investigation. Exploring fine-tuning techniques could enhance its generalizability across diverse text types.

On the other hand, the google/PEGASUS-xsum model obtains the highest BART Score of 0.93 on the XSum dataset. The BART Score is a more recent metric that evaluates the quality of generated text using pre-trained language models like BART. The high BART Score suggests that the summaries generated by the google/PEGASUS-xsum model on the XSum dataset are of high quality and closely resemble human-written summaries.

Comparing the models, it is evident that their performance varies depending on the dataset they are evaluated on. This highlights the importance of choosing the appropriate model for a specific task and dataset. Fine-tuning a pre-trained model like PEGASUS on a dataset similar to the target application can lead to better results.

In Table 4, the R-L score for the Model-Newsroom on the Gigaword dataset is 0.0104. The value is very low.

The R-L score, or R-L, measures the longest common subsequence between the generated summary and the reference summary. A score of 0.0104 suggests that the model's generated summaries have very little overlap with the reference summaries for the Gigaword dataset.

The R-L score of 0.0104 for the google/PEGASUS-Newsroom model on the Gigaword dataset is very low but not exactly zero. This could be attributed to the model's fine-tuning on a different dataset, limitations of the ROUGE metrics, or specific challenges posed by the Gigaword dataset.

In conclusion, the google/PEGASUS-Gigaword model achieves the highest scores for the R-1, R-2, and R-L metrics on the Gigaword dataset, while the google/PEGASUS-xsum model obtains the highest BART Score on the XSum dataset. The performance of these models varies across different datasets, emphasizing the importance of selecting the appropriate model and fine-tuning it for a specific task and dataset. These findings can be applied to other text generation tasks, and the choice of evaluation metric should be carefully considered to ensure a comprehensive assessment of the generated text's quality.

Table 4. FLAN-T5 base performance metrics across diverse datasets

Model (Flan-T5-base)	Dataset	R-1	R-2	R-L	BART Score
summary	XSum	0.2832	0.0857	0.2209	0.89
	CNN/DailyMail	0.1612	0.0465	0.1208	0.85
	Multi-News	0.1336	0.04	0.0817	0.83
	Newsroom	0.1815	0.1131	0.1638	0.85
	Gigaword	0.5587	0.2677	0.5107	0.89
multi_answer	XSum	0.2757	0.0988	0.2267	0.89
	CNN/DailyMail	0.2046	0.0895	0.163	0.86
	Multi-News	0.14696	0.0638	0.0999	0.84
	Newsroom	0.2092	0.1536	0.1965	0.86
	Gigaword	**0.5271**	**0.2716**	**0.5019**	**0.91**

Based on the table, the highest values for each metric across all models and datasets are:

- Highest R-1: 0.5587 (Flan-T5(summary) on Gigaword dataset)
- Highest R-2: 0.2716 (Flan-T5 (summary) on Gigaword dataset)
- Highest R-L: 0. 5107 (Flan-T5 (summary) on Gigaword dataset)
- Highest BART Score: 0.91 (Flan-T5-base (multi_answer) on Gigaword dataset)

Analysis: The Flan-T5-base model, when fine-tuned for summarization (summary), achieves the highest scores for the R-1, R-2, and R-L metrics on the Gigaword dataset. This suggests that the model is well-suited for generating summaries on texts similar to those in the Gigaword dataset. However, its performance varies when applied to other datasets like XSum, CNN/DailyMail, Multi-News, and Newsroom, indicating that the model's performance is dependent on the nature of the input text.

Interestingly, when the Flan-T5-base model is fine-tuned for question answering to generating multiple answers (multi_answer), it achieves the highest BART Score of 0.91 on the Gigaword dataset. The high BART Score indicates that the generated summaries are of high quality and closely resemble human-written summaries.

Comparing the different fine-tuning approaches (summary and multi_answer) for the Flan-T5-base model, we observe that the performance varies slightly across the datasets. This suggests that the choice of fine-tuning approach can impact the model's performance on different tasks and datasets.

When applying these findings to other tasks, it is essential to consider the similarity between the pre-training data, fine-tuning data, and the target task data. Fine-tuning on a dataset that closely resembles the target task data can result in better performance.

Fine-tuning the model on a dataset similar to the target task data can lead to better performance. For instance, the Flan-T5-base model performs best on the Gigaword dataset when fine-tuned for summarization.

4.3. Dataset Selection

The nature of the dataset significantly affects model performance. It is crucial to select datasets that closely match the target application for fine-tuning.

Evaluation Metrics:

The choice of evaluation metrics should align with the specific task. While R-1, R-2, and R-L are suitable for summarization tasks, other metrics like accuracy, F1 score, or task-specific metrics may be more appropriate for different tasks.

Furthermore, the choice of evaluation metrics should be carefully considered based on the specific task. While R-1, R-2, and R-L are commonly used for summarization tasks, other metrics like accuracy, F1 score, or task-specific metrics may be more appropriate for different tasks.

In conclusion, the Flan-T5-base model, when fine-tuned for summarization, achieves the highest scores for the R-1, R-2, and R-L metrics on the Gigaword dataset. When fine-tuned for question answering to generating multiple answers, it obtains the highest BART Score on the same dataset. The performance of the model varies across different datasets and fine-tuning approaches, highlighting the importance of selecting the appropriate fine-tuning strategy and dataset for a specific task. These findings can be applied to other NLP tasks, and the choice of evaluation metrics should be tailored to the specific task at hand.

4.4. Efficiency of Summarization by FLAN-T5 and PEGASUS Models

We evaluated summarization efficiency by comparing generated summary lengths against reference lengths for FLAN-T5 and PEGASUS across five datasets (XSum, CNN/DailyMail, Multi-News, Newsroom, and Gigaword), as illustrated in Figures 5 and 6.

- PEGASUS Model (Figure 5): PEGASUS demonstrated variable efficiency. On XSum, it consistently generated summaries shorter than references (average 17.3-20.8 words). Conversely, on CNN/DailyMail, summaries were longer (26.6-46.2 words). Notably, PEGASUS produced significantly longer summaries on Multi-News (88.8-170.5 words), suggesting challenges in compressing multi-document inputs efficiently.
- FLAN-T5 Model (Figure 6): FLAN-T5 (t5_summary variant) exhibited more consistent summary lengths. For XSum, generated (23.2 words) and reference (21.3 words) lengths were comparable. However, similar to PEGASUS, FLAN-T5 also generated longer summaries for Multi-News (48.3-62.0 words), indicating difficulty with this dataset.

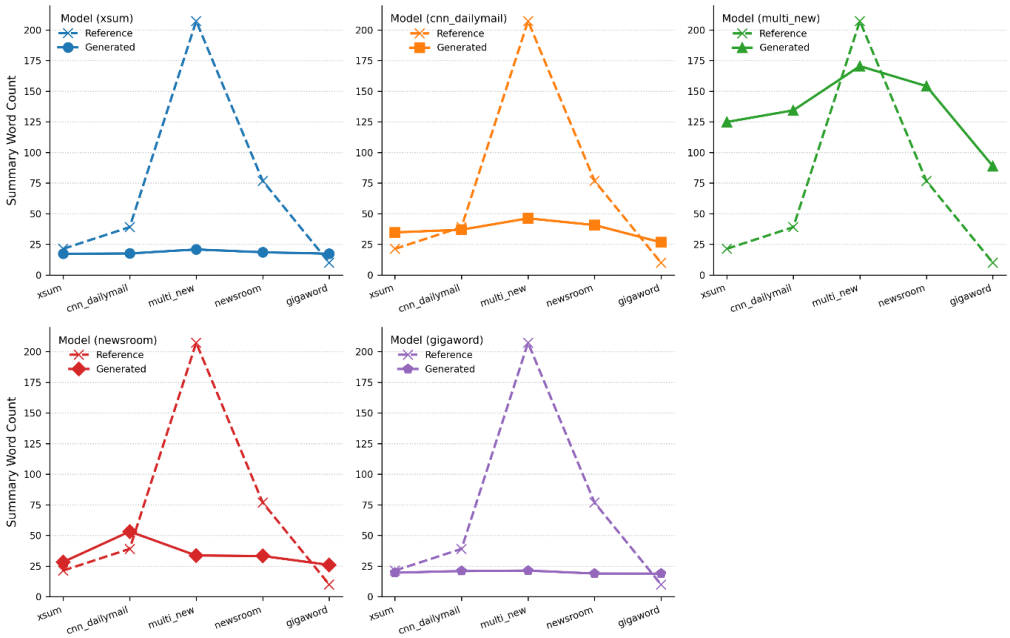

Fig.5. Comparison of generated and reference summary word counts for PEGASUS across datasets

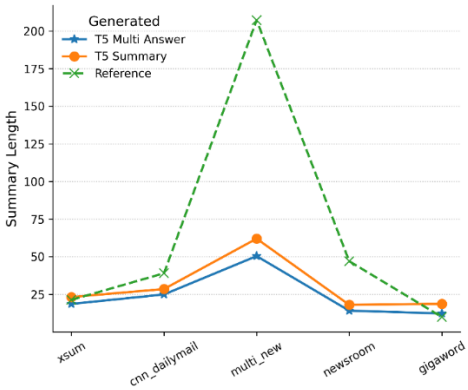

Fig.6. Comparison of generated and reference summary word counts for FLAN-T5 across datasets

4.5. Comparison and Implications for Efficiency

Both PEGASUS and FLAN-T5 effectively compressed original documents, but their efficiency, particularly regarding summary length, varied. FLAN-T5 generally produced more consistent summary lengths across datasets (Figure 6), suggesting better generalization in this aspect. PEGASUS showed greater fluctuations (Figure 5), especially struggling with Multi-News inputs, leading to much longer summaries.

These observed differences in summarization length and efficiency are significantly influenced by model configuration parameters, detailed in Table 1 ('Configuration Details of Summarization Models'). Key parameters include:

- Length Controls: max_length (sets upper summary bound, balancing detail with coherence), min_length (ensures minimum informativeness, avoiding redundancy), and length_penalty (encourages conciseness).
- Processing Capacity & Quality: max_position_embeddings (determines capacity for longer sequences), num_beams (impacts output quality vs. computational cost), d_model (dimensionality affecting information

capture), and vocab_size (word recognition range).

Optimizing these parameters is crucial for tailoring summarization efficiency. For instance, higher max_length allows more detail but can reduce conciseness, while a higher length_penalty promotes brevity. Similarly, larger d_model or num_beams can improve quality but increase computational load. The struggle of both models with Multi-News highlights the challenge of configuring models for optimal efficiency on very long or complex multi-document inputs, necessitating further research in this area. Nonetheless, both models showcase the potential of advanced language models for text summarization.

4.6. Performance Implications

The comprehensive analysis presented in the "Results and Discussion" section offers a nuanced understanding of the operational characteristics, scalability, and complexity of various summarization models, including PEGASUS variants and the FLAN-T5 model. This analysis is pivotal for advancing the field of NLP by elucidating the intricate balance between model performance, operational efficiency, and complexity. The findings underscore the importance of meticulously configuring model parameters to optimize performance, a task that requires a deep understanding of each parameter's impact on the model's capabilities and limitations.

4.7. Operational Characteristics and Model Performance

The operational parameters, such as max position embeddings, max length, min length, length penalty, number of beams, d_model, and vocabulary size, play a critical role in determining a model's performance and efficiency. For instance, max position embeddings are essential for processing extensive documents, but they also increase computational demands. Similarly, the max length parameter influences the detail level of summaries, impacting coherence. These parameters must be carefully balanced to achieve the desired model performance without compromising operational efficiency.

4.8. Scalability and Complexity

The scalability of a model is directly influenced by its configuration parameters. Models with higher max position embeddings and larger d_model values can process more complex information but at the cost of increased computational requirements. This trade-off highlights the challenge of scaling models to handle extensive or complex datasets without incurring prohibitive computational costs. Furthermore, the complexity of a model, as determined by its configuration, affects its ability to generalize across different tasks and datasets. Models with larger vocabularies and more beams in the beam search algorithm demonstrate enhanced word recognition capabilities and output quality but may suffer from slower processing speeds and higher memory requirements.

4.9. Implications for Future Research

The insights derived from this analysis have significant implications for future research in NLP. They highlight the necessity of optimizing model configurations to balance performance, efficiency, and complexity. This optimization is crucial for developing models that can effectively process and summarize extensive or complex texts without excessive computational demands. Additionally, the findings suggest avenues for further research, such as exploring alternative configurations that could enhance model scalability and reduce complexity without compromising performance.

In conclusion, this detailed examination of summarization models' operational characteristics, scalability, and complexity provides valuable insights for the NLP community. It emphasizes the importance of strategic parameter configuration and opens up new research directions aimed at improving model efficiency and performance. This analysis not only advances our understanding of summarization models but also sets the stage for future innovations in the field.

5. Conclusions

This research significantly advances abstractive summarization by providing a comprehensive evaluation of PEGASUS and Flan-T5 across diverse datasets, revealing PEGASUS's strength in generating coherent summaries for large-scale datasets, while establishing Flan-T5, via our novel T5 Dual Summary Integration Framework, as a leader in efficient and nuanced summarization. This dual framework represents a concrete advancement by directly addressing the common trade-off between summary length and depth; it leverages Flan-T5's adaptability to generate both a concise general overview and a detailed, structured summary incorporating multiple information dimensions (e.g., Trends, Facts, Vision, History). This allows for controlled conciseness, producing summaries closely aligned with reference lengths (as shown in Section 4.6), without sacrificing the contextual richness achievable through the structured components. This innovative approach, balancing brevity and accuracy, demonstrates the potential of modular architectures. The ability to provide layered information makes this framework particularly impactful for use cases requiring variable depth, such as business intelligence reporting, scientific literature review, news aggregation, and legal document analysis. Furthermore, the study exposes critical shortcomings in existing evaluation metrics, notably ROUGE’s inadequacy for nuanced assessment and its moderate correlation with BARTScore, which prioritizes surface-level accuracy over semantic richness and contextual fidelity. These limitations highlight the urgent need for hybrid evaluation frameworks integrating

quantitative metrics, semantic analysis, and human input to drive true progress. This need for more robust evaluation will likely be amplified by advancements in hybrid summarization techniques [51] , which combine methods to enhance summary quality but will require more nuanced assessment. Encouragingly, the fact that recent research has focused on developing improved hybrid evaluation metrics, such as HybridEval [52], designed to better capture semantic fidelity, underscores the timeliness of this issue.Moreover, variability in model performance across datasets challenges the notion of universal applicability, emphasizing task-specific optimization for tailored NLP solutions. Future work will pursue several key directions: enhancing the dual summary framework via advanced integration and adaptive learning (e.g., reinforcement learning); improving model robustness by expanding training data to diverse domains (medical, legal, technical) and multilingual/cultural contexts; exploring complementary NLP advancements like retrieval-augmented generation and multimodal summarization to boost summary quality; and investigating methods to enhance model interpretability, acknowledging its importance for understanding and trusting summarization outputs. Comprehensive evaluation using both automated metrics and human assessment will remain crucial across these efforts.

References

[1] K. Yao, L. Zhang, D. Du, T. Luo, L. Tao, and Y. Wu, "Dual encoding for abstractive text summarization," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 985–996, 2020, doi: 10.1109/TCYB.2018.2876317.

[2] A. M. A. Zeyad and A. Biradar, "A Hybrid Text Summarization Approach Using Neural Networks and Metaheuristic Algorithms," *Int. J. Saf. Secur. Eng.*, vol. 13, no. 3, pp. 479–489, 2023, doi: 10.18280/ijsse.130310.

[3] R. Pasunuru and M. Bansal, "Multi-reward reinforced summarization with saliency and entailment," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 2, pp. 646–653, 2018, doi: 10.18653/v1/n18-2102.

[4] W. Kryściński, R. Paulus, C. Xiong, and R. Socher, "Improving abstraction in text summarization," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 1808–1817, 2018, doi: 10.18653/v1/d18-1207.

[5] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization," in *37th International Conference on Machine Learning, ICML 2020*, 2020, pp. 11328–11339. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3524938.3525989

[6] C. Raffel *et al.*, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Bloomsbury Publishing, 2020. doi: 10.5040/9781408182406.00000009.

[7] Y. K. Atri, V. Goyal, and T. Chakraborty, "Multi-Document Summarization Using Selective Attention Span and Reinforcement Learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3457–3467, 2023, doi: 10.1109/TASLP.2023.3316459.

[8] T. Hirao, M. Nishino, J. Suzuki, and M. Nagata, "Enumeration of extractive oracle summaries," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 1, pp. 386–396, 2017, doi: 10.18653/v1/e17-1037.

[9] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 708–719. doi: 10.18653/v1/N18-1065.

[10] A. M. Rush, S. Chopra, and J. Weston, "A Neural Attention Model for Abstractive Sentence Summarization," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 379–389. doi: 10.18653/v1/D15-1044.

[11] M. Zhang, C. Li, M. Wan, X. Zhang, and Q. Zhao, "ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics," *Expert Syst. Appl.*, vol. 237, no. PA, p. 121364, 2024, doi: 10.1016/j.eswa.2023.121364.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating Text Generation With Bert," 2020, doi: https://doi.org/10.48550/arXiv.1904.09675.

[13] A. R. Fabbri *et al.*, "SummEval: Re-evaluating Summarization Evaluation," *Trans. Assoc. Comput. Linguist.*, pp. 391–409, 2021, doi: 10.1162/tacl_a_00373.

[14] G. Moro and L. Ragazzi, "Align-then-abstract representation learning for low-resource summarization," *Neurocomputing*, vol. 548, pp. 1–9, 2023, doi: https://doi.org/10.1016/j.neucom.2023.126356.

[15] M. Sänger, L. Weber, and U. Leser, "WBI at MEDIQA 2021: Summarizing Consumer Health Questions with Generative Transformers," *Assoc. Comput. Linguist.*, pp. 86–95, 2021, doi: 10.18653/v1/2021.bionlp-1.9.

[16] M. Zhang, G. Zhou, W. Yu, N. Huang, and W. Liu, "A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–21, Aug. 2022.

[17] A. G. Etemad, A. I. Abidi, and M. Chhabra, "A Review on Abstractive Text Summarization Using Deep Learning," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1–6. doi: 10.1109/ICRITO51393.2021.9596500.

[18] J. Zhong, Z. Wang, and others, "Mtl-das: Automatic text summarization for domain adaptation," *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[19] 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," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 1797–1807. doi: 10.18653/v1/D18-1206.

[20] A. P. Widyassari, A. Affandy, E. Noersasongko, A. Z. Fanani, A. Syukur, and R. S. Basuki, "Literature Review of Automatic Text Summarization: Research Trend, Dataset and Method," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, 2019, pp. 491–496. doi: 10.1109/ICOIACT46704.2019.8938454.

[21] S. Rothe, J. Maynez, and S. Narayan, "A Thorough Evaluation of Task-Specific Pretraining for Summarization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 140–145. doi: 10.18653/v1/2021.emnlp-main.12.

[22] G. Tsuchiya, "Postmortem Angiographic Studies on the Intercoronary Arterial Anastomoses.: Report I. Studies on Intercoronary Arterial Anastomoses in Adult Human Hearts and the Influence on the Anastomoses of Strictures of the Coronary Arteries.," *Jpn. Circ. J.*, vol. 34, no. 12, pp. 1213–1220, 1971, doi: 10.1253/jcj.34.1213.

[23] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1029–1046, 2022, doi: https://doi.org/10.1016/j.jksuci.2020.05.006.

[24] Y. D. Prabowo, A. I. Kristijantoro, H. Warnars, and W. Budiharto, "Systematic literature review on abstractive text summarization using kitchenham method," *ICIC Express Lett. Part B Appl.*, vol. 12, no. 11, pp. 1075–1080, 2021.

[25] P. Wang *et al.*, "Multi-Document Scientific Summarization from a Knowledge Graph-Centric View," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6222–6233. [Online]. Available: https://aclanthology.org/2022.coling-1.543

[26] A. Cohan, G. Feigenblat, T. Ghosal, and M. Shmueli-Scheuer, "Overview of the First Shared Task on Multi Perspective Scientific Document Summarization ({M}u{P}),," in *Proceedings of the Third Workshop on Scholarly Document Processing*, A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. de Waard, and L. L. Wang, Eds., Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 263–267. [Online]. Available: https://aclanthology.org/2022.sdp-1.32

[27] X. Tie *et al.*, "Personalized Impression Generation for PET Reports Using Large Language Models," *J. Imaging Informatics Med.*, pp. 1–18, 2024.

[28] R. Tang, G. Lueck, R. Quispe, H. Inan, J. Kulkarni, and X. Hu, "Assessing Privacy Risks in Language Models: A Case Study on Summarization Tasks," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15406–15418. doi: 10.18653/v1/2023.findings-emnlp.1029.

[29] S. Ruder, I. Vulić, and A. Søgaard, "Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold," *arXiv Prepr. arXiv2206.09755*, 2022.

[30] C. P. Chai, "Comparison of text preprocessing methods," *Nat. Lang. Eng.*, vol. 29, no. 3, pp. 509–553, 2023.

[31] S. Wadhwa, S. Amir, and B. C. Wallace, "Revisiting relation extraction in the era of large language models," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023, p. 15566.

[32] H. W. Chung *et al.*, "Scaling Instruction-Finetuned Language Models," pp. 1–54, 2022, doi: https://doi.org/10.48550/arXiv.2210.11416.

[33] G. Raposo, L. Coheur, and B. Martins, "Prompting, Retrieval, Training: An exploration of different approaches for task-oriented dialogue generation," *Assoc. Comput. Linguist.*, pp. 400–412, 2023, doi: 10.18653/v1/2023.sigdial-1.37.

[34] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization," in *37th International Conference on Machine Learning, ICML 2020*, 2020, pp. 11328–11339. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3524938.3525989

[35] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, no. Nips, pp. 5999–6009, 2017.

[36] J. W. Rae *et al.*, "Scaling language models: Methods, analysis \& insights from training gopher," *arXiv Prepr. arXiv2112.11446*, 2021.

[37] A. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1074–1084. doi: 10.18653/v1/P19-1102.

[38] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv Prepr. arXiv1808.06226*, 2018.

[39] T. B. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 2020-Decem. 2020.

[40] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Front. energy Res.*, vol. 9, p. 652801, 2021.

[41] A. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1074–1084. doi: 10.18653/v1/P19-1102.

[42] M. S. and P. J. L. Jingqing Zhang, Yao Zhao, "XSum Dataset." [Online]. Available: https://huggingface.co/google/pegasus-xsum

[43] M. S. and P. J. L. Authors: Jingqing Zhang, Yao Zhao, "CNN/DailyMail Dataset." [Online]. Available: https://huggingface.co/google/pegasus-cnn_dailymail

[44] M. S. and P. J. L. Authors: Jingqing Zhang, Yao Zhao, "Multi-News Dataset." [Online]. Available: https://huggingface.co/google/pegasus-multi_news

[45] M. S. and P. J. L. Authors: Jingqing Zhang, Yao Zhao, "Newsroom Dataset." [Online]. Available: https://huggingface.co/google/pegasus-newsroom

[46] M. S. and P. J. L. Authors: Jingqing Zhang, Yao Zhao, "Gigaword Dataset." [Online]. Available: https://huggingface.co/google/pegasus-gigaword

[47] M. Lewis and et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.703/

[48] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Q. Weinberger, "Latent Diffusion for Language Generation." 2023. [Online]. Available: https://arxiv.org/abs/2212.09462

[49] M. T. R. Laskar, M. S. Bari, M. Rahman, M. A. H. Bhuiyan, S. Joty, and J. Huang, "A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 431–469. doi: 10.18653/v1/2023.findings-acl.29.

[50] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive Summarization as Text Matching," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 6197–6208. doi: 10.18653/v1/2020.acl-main.552.
[51] J. Junadhi, A. Agustin, L. Efrizoni, F. Okmayura, D. R. Habibie, and M. Muslim, "Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 885–896, 2025.
[52] R. Sarwar *et al.*, "HybridEval: An Improved Novel Hybrid Metric for Evaluation of Text Summarization," *J. Informatics Web Eng.*, vol. 3, no. 3, pp. 233–255, 2024.

Authors' Profiles

Mr. Abdulrahman Mohsen Ahmed Zeyad is a Ph.D. scholar at the School of Computer Science and Engineering, REVA University. His doctoral research focuses on Natural Language Processing (NLP).

Dr. Arun Biradar, Professor at REVA University's School of Computer Science and Engineering, holds a Ph.D. in CSE and has 25 years of experience. His research spans MANET, Genetic Algorithms, SE, and AI. He is an ISTE awardee, supervises PhD candidates (3 completed/8 ongoing), has 40+ publications, 4 copyrights, guided 65+ projects, and secured 9 Lakhs in funding.

How to cite this paper: Abdulrahman Mohsen Ahmed Zeyad, Arun Biradar, "Abstractive Text Summarization: A Hybrid Evaluation of Integrating Flan-T5 (Dual Framework) with Pegasus Reveals Conciseness Advantages across Diverse Datasets", International Journal of Computer Network and Information Security(IJCNIS), Vol.17, No.6, pp.98-115, 2025. DOI:10.5815/ijcnis.2025.06.07