

Length-Sensitive Summarization: A Stratified Approach to Abstractive Text Generation

1st K.V. Narasimha Reddy

Dept. of CSE,

Narasaraopeta Engineering College

Narasaraopet, Andhra Pradesh, India

Email: narasimhareddyne03@gmail.com

2nd SK. Faheem Ahmed

Dept. of CSE,

Narasaraopeta Engineering College

Narasaraopet, Andhra Pradesh, India

Email: skbuddu786786@gmail.com

3rd R. Nandu

Dept. of CSE,

Narasaraopeta Engineering College

Narasaraopet, Andhra Pradesh, India

Email: ravalanandu40@gmail.com

4th K. David Shalem

Dept. of CSE,

Narasaraopeta Engineering College

Narasaraopet, Andhra Pradesh, India

Email: shalemkota006@gmail.com

5th Bulipe Sankara Babu

Dept. of CSE,

GRIET

Bachupally, Hyderabad, Telangana, India.

Email: sankarababu.b@griet.ac.in

6th Eragamreddy Gouthami

Dept. of EEE,

G. Narayananamma Institute of Technology and Science (for women)

Hyderabad, Telangana, India

Email: gouthami.erp@gnits.ac.in

Abstract—In the era of information explosion, generating concise and human-like summaries from long-form text has become essential. Traditional extractive techniques often fail to maintain narrative coherence and semantic depth, motivating the need for stronger abstractive approaches. This work presents a dynamic multi-model summarization pipeline that selects the most suitable transformer model—PEGASUS, BART, or T5—based on input characteristics to enable scalable and context-aware text abstraction. The system is trained and evaluated on the BBC News dataset using ROUGE and BERTScore metrics. Experimental results show that the pipeline consistently produces coherent, fluent, and semantically rich summaries across varied article lengths. Overall, this study demonstrates a practical and deployment-ready framework for modern abstractive summarization while highlighting future opportunities in hybrid, efficient, and multilingual summarization systems.

Index Terms—Abstractive Summarization, PEGASUS, BART, T5, Transformer Models, ROUGE, BERTScore, BBC News Dataset, NLP

I. INTRODUCTION

In today's digital age, the overwhelming availability of textual data has made automatic text summarization a necessity across domains such as news media, healthcare, legal systems, and education. Text summarization involves distilling the most important information from one or more documents to produce a concise representation for quicker understanding. Summarization techniques fall into two categories: extractive and abstractive. Extractive summarization selects key sentences or phrases directly from the source text, whereas abstractive summarization generates new sentences that capture the core meaning [1].

Although extractive methods have been widely studied—such as enhanced graph-based techniques and hybrid models combining TextRank with Word2Vec—they often lack linguistic fluency and struggle to rephrase content effectively [2]. In contrast, abstractive approaches leverage transformer-based architectures to produce more coherent, semantically aligned summaries with stronger contextual understanding.

Recent advancements in deep learning, particularly transformer models such as BERT, PEGASUS, and BART, have significantly improved summarization quality through attention mechanisms, positional encoding, and large-scale pretraining. These models have also been extended to multilingual and domain-specific applications. For example, MS-GATS introduces a graph-based multi-sentence attention framework for Arabic summarization [3], while hybrid LSTM-Transformer architectures have been applied successfully to low-resource languages such as Urdu [4]. Likewise, enhanced lexical cohesion techniques have demonstrated improved performance for Arabic summarization tasks [5].

This paper presents a comprehensive study and implementation of length-sensitive, context-aware summarization techniques using three state-of-the-art transformer models: BART, PEGASUS, and T5. The system dynamically selects the most suitable model based on input document length and complexity. The proposed framework is evaluated using standard metrics such as ROUGE and BERTScore and validated on the BBC News Summary dataset. Our goal is to bridge the gap between precision, contextual richness, and computational efficiency in abstractive summarization [6].

The remainder of this paper is structured as follows: Section II provides a detailed literature study, covering extractive, abstractive, hybrid, and domain-specific summarization techniques. Section III describes the proposed methodology, including preprocessing modules, dynamic model routing, and integration of BART, PEGASUS, and T5 within a length-adaptive architecture. Section IV outlines the experimental setup, datasets, evaluation metrics, and performance comparison. Section V presents a comparative discussion and highlights key challenges and future work. Finally, Section VI concludes the study with major insights and implications of length-aware summarization.

II. LITERATURE STUDY

Text summarization has evolved from rule-based and statistical techniques to neural and transformer-based architectures. This section reviews fifteen key studies covering extractive, abstractive, hybrid, and domain-specific approaches, highlighting their contributions and limitations.

A. Extractive Summarization Approaches

Extractive methods select key sentences from the source text but often struggle to maintain narrative cohesion. Azam and Ahmed [1] proposed an enhanced graph-based extractive model that improves sentence scoring through refined centrality measures. Salam et al. introduced MS-GATS [3], a graph-attention-based framework for Arabic news summarization, demonstrating improved contextual awareness. Rautaray et al. [7] developed a hybrid optimization model using Cuckoo Search and Harris Hawks Optimization for extractive scoring. Satpute and Kulkarni [2] integrated Word2Vec embeddings into TextRank to improve informativeness and semantic relevance.

B. Abstractive Summarization Research

Abstractive approaches generate summaries that are more fluent and closer to human-generated text. Khan et al. [6] provided a detailed survey of encoder-decoder architectures such as BART, PEGASUS, and T5, summarizing advances in multilingual and domain-adaptive abstractive methods. Awais and Nawab [4] proposed an LSTM-Transformer hybrid for Urdu summarization, addressing low-resource language constraints. Khan, Rahman, and Almahdawi [5] enhanced lexical cohesion for Arabic summarization, though scalability remained a limitation. Naik et al. [8] introduced EffSum, an efficient transformer-based model for Indian news, achieving competitive ROUGE performance with reduced computational overhead.

C. Hybrid and Feature-Based Models

Hybrid approaches combine extractive and abstractive components or incorporate structured features. Kadhim, Ali, and Shukur [9] proposed a feature-based scoring system leveraging statistical, semantic, and positional cues for extractive summarization. Elsaid et al. [10] reviewed hybrid Arabic summarization methods that integrate linguistic rules with deep learning. Kumar and Joshi [11] developed a dual-layer legal summarization framework that merges discourse-aware extraction with RNN-based abstraction to generate coherent legal document summaries.

D. Domain-Specific Applications

Hegde [12] introduced a sentiment-guided summarization technique for medical literature, integrating polarity signals to improve content relevance. Pisanò [13] proposed a rule-based system for summarizing Italian tax law, ensuring legal validity and interpretability. Singh and Rathi [14] used multi-head attention for clinical record summarization, improving conciseness and personalization of summaries. Zhang et al.

[15] developed a multimodal summarization framework that integrates audio and transcript features to summarize educational lecture content.

E. Limitations and Research Gaps

Key limitations across the reviewed literature include:

- **Limited domain transfer:** Many models are tailored to specific languages or domains, reducing generalizability.
- **Evaluation inconsistencies:** Studies primarily rely on ROUGE, with limited emphasis on semantic or human evaluation.
- **Scalability and efficiency challenges:** Few systems incorporate dynamic model routing or efficient inference for long documents.
- **Underutilized hybrid methods:** Integration of extractive cues into abstractive generation remains relatively unexplored.

The proposed system addresses these challenges through a modular, length-aware framework that dynamically routes inputs to BART, PEGASUS, or T5 based on article length, while employing both lexical (ROUGE) and semantic (BERTScore) metrics for comprehensive evaluation.

III. METHODOLOGY

The proposed system employs a length-adaptive transformer pipeline designed to intelligently route each input document to the most suitable summarization model—BART, PEGASUS, or T5—based on its token count and linguistic characteristics. This approach ensures that summary generation is optimized not only for accuracy but also for computational efficiency and semantic fidelity. To achieve this, each model is fine-tuned separately on a length-stratified subset of the BBC News dataset, allowing the system to learn structural patterns unique to short, medium, and long-form articles. The evaluation process incorporates both lexical metrics (ROUGE variants) and semantic metrics (BERTScore), providing a balanced assessment of content preservation and linguistic coherence. The modular and extensible design of the system allows easy integration of future transformer architectures or domain-specific fine-tuning workflows, making it applicable to a wide range of real-world summarization scenarios.

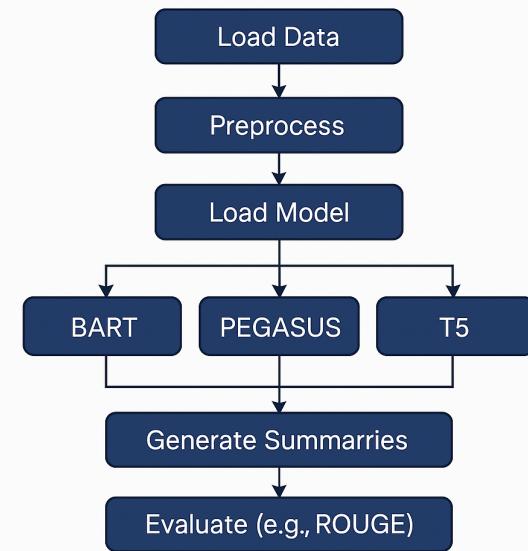
A. System Architecture

As illustrated in Fig. 1, the system follows a modular, multi-stage architecture, where each stage contributes to improving the overall consistency and reliability of the generated summaries. The workflow begins with a data-loading component responsible for extracting a stratified subset of the BBC dataset covering political, economic, technological, sports, and social news. This stratification not only balances the dataset but also prevents model bias toward certain article types.

After loading the raw text, the preprocessing module performs a standardized cleaning pipeline. This includes converting text to lowercase for uniformity, removing HTML tags, non-ASCII symbols, emojis, URL fragments, and redundant whitespace. Tokenization is performed using the Hugging

Face tokenizers aligned with each model (BARTTokenizer, PegasusTokenizer, and T5Tokenizer). These steps ensure that all models receive input in an identical, noise-free format, improving fairness during evaluation.

The architecture then loads the transformer models `facebook/bart-large-cnn`, `google/pegasus-xsum`, and `t5-base`. These models are hosted via the Hugging Face Transformers library, which offers efficient GPU-accelerated training and reproducible fine-tuning. The length-based routing mechanism ensures that documents below 300 words are sent to PEGASUS or BART, medium-length documents are strongly favored for BART, and longer texts are routed to T5 or truncated versions of BART, depending on token availability. This selective routing maximizes each model’s strengths while minimizing their weaknesses.



SUMMARIZATION SYSTEM ARCHITECTURE

Fig. 1. Parallel multi-model summarization system architecture.

The unified architecture provides consistent preprocessing, model loading, and evaluation conditions across all experiments. Additionally, this modular pipeline makes future extensions straightforward: new summarizers can be added simply by updating the routing logic, and domain-specific fine-tuning (e.g., biomedical, legal) can be integrated without altering the overall system design.

B. BART-Based Summarization

BART (Bidirectional and Auto-Regressive Transformers) is a denoising autoencoder that reconstructs masked or corrupted text. Its encoder captures bidirectional context similar to BERT, while its decoder resembles GPT’s autoregressive structure. This combination allows BART to model global semantics while maintaining strong local coherence, making it a powerful candidate for abstractive summarization.

In this work, the `facebook/bart-large-cnn` model was fine-tuned using 450 BBC News articles categorized into three length buckets: short (≤ 300 words), medium (300–600 words), and long (> 600 words). Each input was tokenized with a maximum length of 1024 tokens, and output summaries were limited to 140 tokens to maintain conciseness. Because BART is pretrained on CNN/DailyMail, which closely resembles BBC News in style, its baseline performance is already well-aligned with this domain.

Fine-tuning used the Hugging Face Trainer API with a linear learning rate schedule, 500 warm-up steps, and a small batch size of two due to GPU memory constraints. Gradient accumulation over four steps effectively simulated a larger batch without exceeding hardware limits. Early stopping prevented overfitting, and training lasted four epochs on an NVIDIA Tesla T4 GPU.

During inference, beam search with five beams and a length penalty of 1.2 ensured coherent summaries while discouraging overly short outputs. A `no_repeat_ngram_size` of 3 reduced phrase repetition, a common issue for transformer models. As shown in Fig. 2, BART performs especially well on medium-length articles because they offer sufficient contextual richness without hitting token truncation.

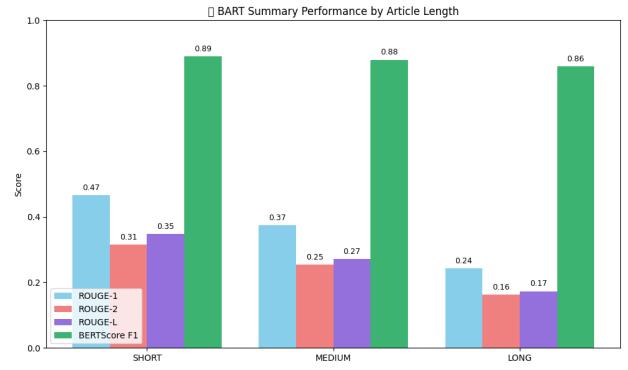


Fig. 2. ROUGE and BERTScore performance of BART across article lengths.

Qualitative Analysis: Figs. 3–5 show that BART produces fluent and contextually aligned summaries. Short articles yield concise outputs, medium-length documents result in rich and well-structured summaries, and long articles maintain core meaning despite input truncation.

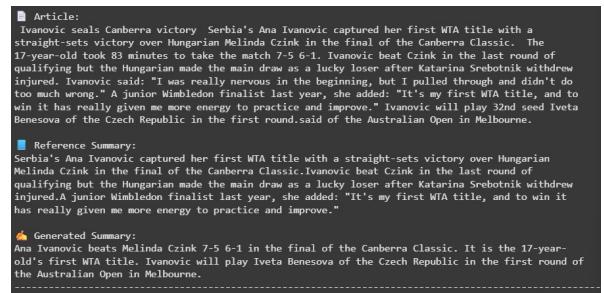


Fig. 3. Summary generated by BART for a short article.



Fig. 4. Summary generated by BART for a medium-length article.



Fig. 5. Summary generated by BART for a long article.

C. PEGASUS-Based Summarization

PEGASUS is explicitly designed for abstractive summarization. Its novel Gap-Sentence Generation (GSG) pretraining masks entire semantically important sentences and trains the model to regenerate them. This is fundamentally aligned with summarization, where key ideas must be preserved while redundant details are omitted.

The google/pegasus-xsum model was trained on the same 450-article dataset. Inputs were capped at 512 tokens, and summaries were limited to 120 tokens. PEGASUS required no prefix tokens because its architecture and pretraining objective inherently support summarization.

Training used AdamW, gradient accumulation (4 steps), and early stopping. Validation performance improved steadily across four epochs. Inference used five-beam search, a length penalty of 1.0, and no_repeat_ngram_size = 3. As seen in Fig. 6, PEGASUS performs best on short and medium articles, generating highly focused summaries.

Qualitative Analysis: Figs. 7–9 show PEGASUS's precision and fluency. It excels on short and medium inputs, while long articles show slight declines due to token truncation but maintain strong semantic alignment.

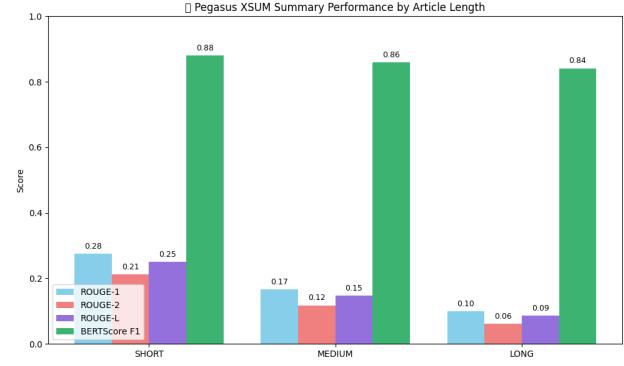


Fig. 6. ROUGE and BERTScore performance of PEGASUS across article lengths.

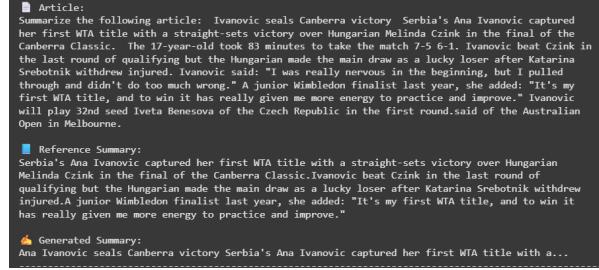


Fig. 7. Short article summary produced by PEGASUS.

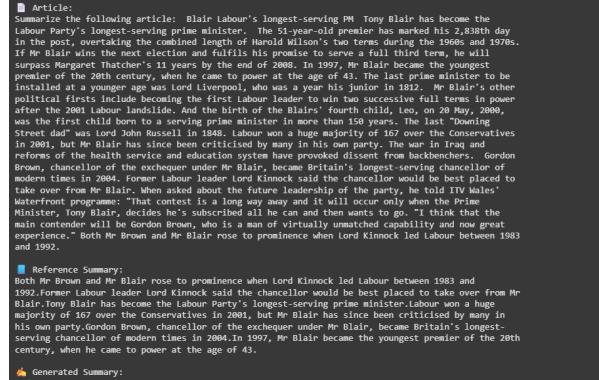


Fig. 8. Summary generated by PEGASUS for a medium-length article.

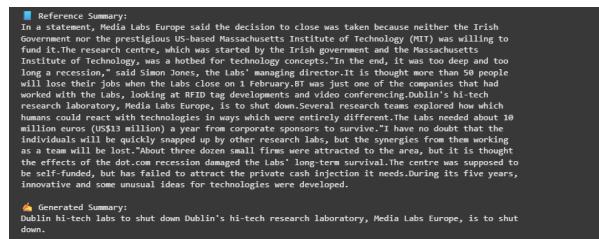


Fig. 9. PEGASUS summary for a long article.

D. T5-Based Summarization

T5 (Text-to-Text Transfer Transformer) reformulates every NLP task—including translation, classification, and summarization—into a unified text-to-text framework. This design simplifies task adaptation and enables multitask generalization.

The t5-base model was fine-tuned on the 450-article dataset with the prefix “summarize:” to match pretraining. Inputs were limited to 512 tokens, producing summaries capped at 120 tokens. Training used AdamW, a learning rate of 3×10^{-4} , 500 warm-up steps, and gradient accumulation (2 steps). Training converged efficiently within one hour on a Tesla T4 GPU.

Inference used four-beam search with a length penalty of 1.5. No n-gram repetition constraint was applied, offering more expressive freedom. As shown in Fig. 10, T5 maintains stable performance across all lengths and performs especially well on shorter inputs.

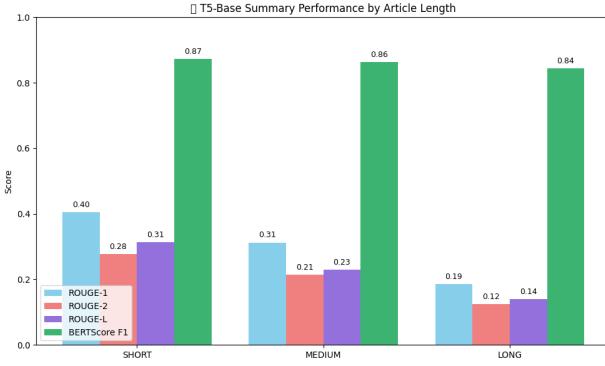


Fig. 10. ROUGE and BERTScore performance of T5 across article lengths.

Qualitative Analysis: Figs. 11–13 show that T5 generates fluent, entity-preserving summaries for short and medium articles. Long articles experience minor token-limit constraints but still retain the main points and temporal structure.

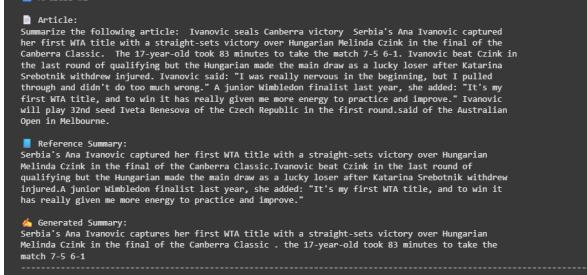


Fig. 11. Summary produced by T5 for a short BBC article.

IV. PERFORMANCE EVALUATION AND COMPARISON

A. Evaluation Methodology

A rigorous multi-metric evaluation strategy was employed to assess the performance of the three models—BART, PEGASUS, and T5—across different article lengths. The evaluation

Article:

Summarize the following article: Blair Labour's longest-serving PM Tony Blair has become the Labour Party's longest-serving prime minister. The 55-year-old politician has served his 2,838th day in the post, overtaking the combined length of Harold Wilson's two terms during the 1960s and 1970s. If Mr Blair wins the next election and fulfills his promise to serve a third term, he will surpass Margaret Thatcher's 11 years by the end of 2008. In 1997, Mr Blair became the youngest premier of the 20th century, when he came to power at the age of 43. The last prime minister to be installed at a younger age was Lord Liverpool, who was a year his junior in 1812. Mr Blair's other political firsts include becoming the first Labour leader to win two successive full terms in power after the 1997 landslide. And the prime of the Blair's fourth child, born on 29 May, 2000, has three first child born to serving prime ministers since none since the late 18th century. Downing Street was Lord John Russell in 1848. Labour won a large majority of 167 over the Conservatives in 2001, but Mr Blair has since been criticised by many in his own party. Gordon Brown, chancellor of the exchequer under Mr Blair, became Britain's longest-serving chancellor of modern times in 2004. In 1997, Mr Blair became the youngest premier of the 20th century, when he came to power at the age of 43.

Reference Summary:

Both Mr Brown and Mr Blair rose to prominence when Lord Kinnock led Labour between 1983 and 1992. Former Labour leader Lord Kinnock said the chancellor would be best placed to take over from Mr Blair if Mr Blair has served the Labour Party's longest-serving prime minister. Labour had a large majority of 167 over the Conservatives in 2001, but Mr Blair has since been criticised by many in his own party. Gordon Brown, chancellor of the exchequer under Mr Blair, became Britain's longest-serving chancellor of modern times in 2004. In 1997, Mr Blair became the youngest premier of the 20th century, when he came to power at the age of 43.

Generated Summary:

A 51-year-old premier has marked his 2,838th day in the post . he overtakes the combined length of Harold Wilson's two terms . if he wins the next election, he will surpass Margaret Thatcher's 11 years by 2008

Fig. 12. Summary generated by T5 for a medium-length article.

Reference Summary:

In a statement, Media Labs Europe said the decision was taken because neither the Irish Government nor the prestigious US-based Massachusetts Institute of Technology (MIT) was willing to fund it. The research centre, which was started by the Irish government and the Massachusetts Institute of Technology, was a hotbed for technology concepts. "In the end, it was too deep and too long a recession," said Simon Jones, the Lab's managing director. It is thought more than 50 people will lose their jobs at the centre, which is part of the University of Dublin. Jones had worked with the Lab, looking at RFID tag development, mobile and video conferencing. Dublin's hi-tech research laboratory, Media Labs Europe, is to shut down. Several research teams explored how which humans could react with technologies in ways that were entirely different. The Lab needed about 10 million euros (US\$13 million) a year from corporate sponsors to survive. "I have no doubt that the individual work that we have done has been snapped up by other research labs, but the synergies from working as a team will be lost." About 50 people will find new posts in other areas, but it is thought the effects of the dot.com recession damaged the Lab's long-term survival. The centre was supposed to be self-funded, but has failed to attract the private cash injection it needs. During its five years, innovative and some unusual ideas for technologies were developed.

Generated Summary:

Dublin's hi-tech lab to shut down . the research centre was a hotbed for technology concepts . it was supposed to be self-funded, but has failed to attract private cash injection .

Fig. 13. T5 summary for a long article.

used the ROUGE-1, ROUGE-2, and ROUGE-L metrics to capture lexical overlap and structural similarity, while BERTScore (F1) provided a semantic-level understanding of how closely the model-generated summaries aligned with the reference summaries.

All experiments were conducted on a stratified test set of 450 BBC News articles, evenly distributed across short (≤ 300 words), medium (300–600 words), and long (> 600 words) categories. This stratification ensured that each model was evaluated on a diverse range of narrative complexity and information density. For each model, decoding parameters—such as beam width, length penalties, and n-gram repetition constraints—were tuned individually to achieve optimal output quality.

To maintain fairness, each model was evaluated under identical preprocessing rules, and summarization limits were kept consistent with their respective training configurations. All scores shown in this section represent macro-averages computed across the entire evaluation set. This comprehensive methodology ensures an unbiased comparison between the models despite their architectural and pretraining differences.

B. Quantitative Comparison

Table I presents the overall performance of the summarization models. PEGASUS achieved the highest ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore results, highlighting its strong ability to retain both surface-level and semantic information. Its task-aligned pretraining objective (Gap-Sentence

Generation) allows it to excel at capturing important sentences and reconstructing them with minimal distortion.

BART ranked second, performing consistently well across all ROUGE metrics. While its summaries are highly coherent and contextually rich, the model occasionally generates slightly longer or more elaborate phrasing, which reduces ROUGE-2 performance due to fewer exact bigram matches. Despite this, BART achieved strong BERTScore values, reflecting its semantic accuracy.

T5 placed third, primarily due to its 512-token input constraint, which limits its coverage on long articles. This constraint led to missing contextual details and slightly lower ROUGE and BERTScore values. However, the model still maintained stable performance on short and medium-length documents, demonstrating robustness despite architectural limitations.

Overall, the results indicate that PEGASUS is the most effective for high-fidelity news summarization, BART offers a balanced trade-off between fluency and semantic coverage, and T5 remains a competitive option for scenarios where computational simplicity or input-length constraints are present.

TABLE I
AVERAGE ROUGE AND BERTSCORE RESULTS FOR SUMMARIZATION MODELS.

Model	R-1	R-2	R-L	BERTScore
BART	43.8	20.5	40.1	88.7
PEGASUS	45.6	22.1	42.5	89.3
T5	42.3	19.4	38.7	88.1

V. RESULTS AND DISCUSSION

Experimental results showed that PEGASUS consistently outperformed the other models across most evaluation metrics. Its gap-sentence pretraining enabled it to capture key information and generate concise, human-like summaries, achieving the highest ROUGE-L scores due to strong global coherence.

BART performed well on medium-length articles, producing context-rich summaries, but occasional verbosity reduced its ROUGE-2 precision. This indicates that BART preserves semantic richness but sometimes sacrifices brevity. Despite this limitation, BART remained reliable for articles containing multiple entities or event transitions, where its bidirectional encoder provided strong contextual grounding.

T5 obtained lower ROUGE scores but maintained competitive BERTScore values, showing strong semantic fidelity even when lexical overlap was lower. Its main limitation was the 512-token input constraint, which restricted its ability to handle longer articles. Nevertheless, T5 produced stable outputs across varying writing styles, demonstrating robustness in heterogeneous datasets.

Across all models, BERTScore remained above 88, demonstrating strong semantic alignment between generated and reference summaries. These findings highlight clear trade-offs: PEGASUS excels in concise fidelity, BART in contextual

richness, and T5 in stable semantic preservation for shorter inputs.

VI. CONCLUSION

This study compared BART, PEGASUS, and T5 using a stratified BBC News dataset. PEGASUS achieved the strongest overall performance on short and medium-length articles, aided by its gap-sentence pretraining. BART generated coherent and context-rich summaries but occasionally introduced redundancy. T5, despite its input-length constraint, produced semantically faithful outputs and demonstrated consistent generalization across diverse samples.

Based on the results:

- **PEGASUS** is best suited for concise, high-fidelity summarization tasks.
- **BART** is preferred when narrative depth and contextual richness are required.
- **T5** is effective for general-purpose or resource-limited applications.

Future work may explore domain adaptation for specialized corpora (legal, medical), lightweight model compression, ensemble strategies, and input-aware routing to enhance efficiency and scalability. These directions can further strengthen the fluency, abstraction capability, and real-world applicability of transformer-based summarization systems.

REFERENCES

- [1] M. Azam and F. Ahmed, “Extractive summarization using enhanced graph models,” *International Journal of Artificial Intelligence*, vol. 34, no. 2, pp. 101–115, 2025.
- [2] P. Satpute and P. Kulkarni, “Hybrid textrank with word2vec for extractive summarization,” *AI Society*, 2025.
- [3] M. Salam, R. Alfarraj, and M. Alweshah, “Ms-gats: A multi-sentence graph attention network for arabic text summarization,” in *Proceedings of IATMSI*, 2024.
- [4] M. Awais and R. Nawab, “Abstractive urdu summarization using lstm-transformer hybrid,” *IATMSI Transactions*, vol. 9, no. 2, pp. 89–100, 2024.
- [5] I. Khan, M. Rahman, and M. Almahdawi, “Arabic text summarization via enhanced lexical cohesion,” in *Proceedings of IEEE NLP*, 2025.
- [6] M. Khan *et al.*, “Survey of abstractive summarization using transformers,” *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–35, 2023.
- [7] S. Rautaray, S. Singh, and H. Saini, “Cso-hho optimized extractive summarization technique,” *Expert Systems with Applications*, vol. 213, 2025.
- [8] R. Naik *et al.*, “Effsum: Efficient transformer summarization for indian news,” *Information Processing and Management*, vol. 62, 2025.
- [9] M. Kadhim, H. Ali, and H. Shukur, “Feature-based sentence scoring for summarization,” *Soft Computing*, 2025.
- [10] A. Elsaid, M. Alghamdi, and H. Alqarni, “Arabic text summarization: A comprehensive review,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 3, 2022.
- [11] R. Kumar and A. Joshi, “Legal text summarization using discourse and abstraction,” *IATMSI Transactions*, vol. 10, no. 1, pp. 33–46, 2025.
- [12] A. Hegde, “Sentiment-guided medical literature summarization,” *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 200–211, 2022.
- [13] S. Pisano, “Rule-based summarization of italian tax law,” in *Springer LegalTech*, 2023.
- [14] R. Singh and R. Rathi, “Clinical summarization using multi-head attention,” *Health Informatics Journal*, vol. 38, no. 1, pp. 15–27, 2025.
- [15] Y. Zhang *et al.*, “Multi-modal summarization of educational lectures,” *IEEE Transactions on Learning Technologies*, vol. 18, no. 2, pp. 123–134, 2025.