ASJET

Al-Salam Journal for Engineering and Technology
Journal Homepage:
http://journal.alsalam.edu.iq/index.php/ajest
e-ISSN: 2790-4822 p-ISSN: 2958-0862

A Comparative Study of Transformer-based and Hybrid Deep Learning Models for Long Document Summarization of academic research papers

Noor Q. Habban[1] * and Mohammed H. Abdulameer[2]

[1]Department of Computer Science Faculty of Computer Science, University of Kufa Najaf, Iraq.
[2]Department of Computer Science Faculty of of Education for Girls University of Kufa Najaf, Iraq.

*Corresponding Author: Noor Q. Habban

DOI: https://doi.org/10.55145/ajest.2025.04.02.005
Received June 2025; Accepted August 2025; Available online August 2025

ABSTRACT: Automatic summarization of long scientific texts has been established as an essential task within the domain of Natural Language Processing(NLP), aiming to reduce information overload and facilitate knowledge acquisition from complex academic documents. Despite its importance, fact-based conventional systems of Automatic Text Summarization have mostly failed in ensuring coherence on the document level and keeping factual correctness. To address these limitations, further recent progress in deep learning has turned more intelligent models toward the equilibrium structural fidelity with fluency. This paper presents a comparative study of two state-of-the-art summarization techniques: (1) hybrid deep learning, which combines Bidirectional LSTM-based sentence classification with a hierarchical attention-driven encoder-decoder for abstractive summarization, and (2) fine-tuned Transformer-based architectures, specifically T5-base and T5-large models. The first method emphasizes structural awareness and hierarchical processing to preserve document-level semantics and mitigate issues such as repetition and out-of-vocabulary (OOV) tokens. In contrast, the Transformer-based models leverage large-scale pretraining with self-attention mechanisms for producing fluent summaries that are richly filled with context. Both methods were evaluated on two benchmark datasets: arXiv and PubMed. The hybrid model achieved ROUGE-F1 scores of (46.7, 19.4, 35.4) and (47.0, 21.3, 39.7), respectively, while the T5-large model outperformed it with scores of (55.8, 33.8, 47.9) and (54.9, 32.0, 48.7). These results show that while Transformer models perform better in abstraction and fluency, the hybrid model has fact control ability that is much interpretable and aligns with document structure. This comparison gives important insight into the trade-off between structure-aware hybrid frameworks and large-scale generative models in academic summarization.

Keywords: extractive summarization, abstractive summarization, hierarchical attention, transformer fine-tuning , Bidirectional LSTM ,T5

1. INTRODUCTION

With the exponential growth of digital text, particularly in scientific and academic fields, the need for efficient and scalable text summarization systems has become increasingly crucial. Automatic Text Summarization (ATS) aims to distill lengthy documents into concise representations that retain core information, thereby aiding information retrieval, literature review, and decision-making [1]. Traditional ATS methods are categorized into three main approaches: extractive, abstractive, and hybrid. Extractive methods rely on selecting salient sentences from the original text [2], while abstractive methods generate new sentences to summarize content humanely [3]. Hybrid models combine the characteristics of both extractive and abstractive methods, often leveraging deep learning techniques to balance factual accuracy with fluency in generated summaries [4].

Recent advancements in deep learning have led to the development of powerful models capable of handling complicated language understanding tasks such as the long-document summarization task. Among these, two fundamentally different yet promising paradigms have emerged: hybrid models that combine bidirectional recurrent networks and attention-based encoders [5], and transformer-based models pre-trained on large corpora [6]. The former makes use of Bidirectional Long Short-Term Memory (BiLSTM) networks for extractive sentence classification and

*Corresponding author: noorqassimh22@gmail.com

hierarchical encoder-decoder architectures that make use of multi-level attention in generating context-aware summaries. The proposed structurally-informed framework is highly effective in ensuring that the semantic coherence is maintained as well as accurately modeling the hierarchical structure of academic literature that is important for ensuring contextually accurate and logically descriptive summaries are produced. In contrast, transformer-based models such as T5-base and T5-large adopt an encoder-decoder framework, incorporating self-attention mechanisms to produce fluent abstract summarization in natural language from the source text. Domain-specific fine-tuning abstracts strong generalization capabilities toward datasets but is limited by compute resources and performance issues for very long sequences of text [7], though they often require substantial computational resources and face limitations when dealing with extremely long input sequences. This study presents a comparative evaluation between these two paradigms, our proposed hybrid BiLSTM + Hierarchical Attention model and transformer-based models (T5-base and T5-large) [8] using benchmark datasets arXiv and PubMed. Evaluation metrics such as ROUGE-1, ROUGE-2, and ROUGE-L are used to assess each model's strengths in terms of content coverage, fluency, and structural coherence. One possible approach is to pre-train the entire model on longer sequences; however, this demands substantial computational resources[9]. Although BERT has shown promise in text summarization, it faces notable limitations in tasks that involve reasoning over long documents[10]. In 2016, the COPYNET model was a sequence-to-sequence framework with an integrated copying mechanism for text summarization. The model was trained on the LCSTS dataset and achieved ROUGE-1: 35.0, ROUGE-2: 22.3, ROUGE-L: 32.0. It effectively handled rare word issues through copying but added complexity during training. The input format was sequential text, and the proposed method utilized Bi-GRU with attention and a copy mechanism[11].

The findings highlight the complementary nature of these models, with the hybrid approach excelling in factual grounding and hierarchical representation, while the transformer models outperform in generating fluent and abstract summaries. The goal of this work is to provide empirical insights into the practical trade-offs between these architectures and to inform future developments in summarization technologies for long scientific texts.

The remainder of this paper is structured as follows: Section 2 reviews the related work and previous studies in academic text summarization. Section 3 describes the proposed hybrid and transformer-based architectures along with the adopted methodology. Section 4 outlines the experimental setup. Section 5 presents the results and a detailed comparative analysis of the models. Finally, Section 6 concludes the paper with important discoveries and suggestions for further study.

2. RELATED WORK

Automatic summarization of long academic texts has given rise to two primary modeling paradigms: Transformer-based approaches and hybrid architectures combining extractive and abstractive techniques. While Transformer models leverage large-scale pretraining and attention mechanisms for fluent and cohesive summaries, hybrid models aim to balance structural integrity with semantic abstraction by integrating sentence selection and hierarchical generation. The following review organizes key contributions thematically to highlight the evolution of techniques and their relevance to long-document summarization tasks.

2.1 Transformer-Based Summarization Models

Transformer models have been extensively used for abstractive summarization due to their capacity for global attention and semantic representation. Also in 2021, researchers fine-tuned the T5 transformer model on XSum and Gigaword datasets, achieving strong performance on short, single-topic news articles. However, the method became less accurate when it had to deal with very long texts, since it was too short in comparison aato the reference texts [12]. And in (2022), BERT-large (as an extractive model) and T5-small (as an abstractive model) were evaluated on data from WikiHow. Higher ROUGE was achieved by BERT, but it did not process data as abstractly as T5, which scored lower due to its size [13]. In another study (2022), PEGASUS-X was introduced for long-document summarization, employing global-local attention and staggered blocks to process inputs up to 16K tokens. Despite strong ROUGE results on arXiv and PubMed, it requires extensive pretraining and large computational resources. Our model, by contrast, achieves effective summarization with fewer parameters using BiLSTM filtering and hierarchical decoding [14]. In 2024, introduced a hybrid model integrating BIGBIRD and DistilBART, supplemented by heuristic sentence scoring for long-document summarization. This approach worked well, but it became difficult to use outside the domains it was trained in due to being based heavily on pretrained models and meta-heuristics algorithms [15].

2.2 Hybrid-Based Summarization Models

Hybrid summarization models combine extractive and abstractive techniques to leverage both sentence-level precision and high-level abstraction. These models are particularly suitable for long and complex academic documents due to their hierarchical nature and modular processing. In 2016, Gu et al. introduced COPYNET, a Seq2Seq model with a copying mechanism that improved handling of OOV words and achieved a ROUGE-1 score of 35.0% for the LCSTS data set. Effectiveness with short texts is there, but using it for long or multilingual texts is still difficult [11]. In 2020, a new hybrid summarization method was developed that combines ARTM topic modeling, structured graphs, and rhetorical structure theory for technical Russian texts. It achieved an ROUGE F1-score of 34.47% and expert-rated

precision of 86.43%, but performance was affected by language complexity and manual setup requirements [16]. In 2021, introduced hybrid summarization models combining BiLSTM, attention, pointer networks, and coverage mechanisms. Their best model (DA-PN + Cover + MLO) improved ROUGE scores on Chinese datasets and reduced common issues like repetition and OOV words. However, it still suffers from shallow architecture and limited use of semantic features [5]. In 2021, used an improved version of TextRank along with a Seq2Seq framework to create summaries for short texts. While effective for brief inputs, because it heavily relied on statistics and its setup did not include a hierarchy, it struggled to process extended and detailed scientific texts [17]. In 2023, developed a BERT-BiGRU model to extract extractive information from long articles in the science documents. For chunk-level features, it applied BERT and used BiGRU with attention to choose the sentences. Tested on arXiv and PubMed, it outperformed baseline models in ROUGE scores, though it suffered from high computational cost and scalability limitations[18]. In 2023, Gurusamy et al. proposed a hybrid model combining Semantic LDA-based extractive summarization with T5-based abstraction. It achieved 48.35%, 29.53%, and 41.72% on DUC2002. However, the method's reliance on concept extraction limits its adaptability to highly varied or very long documents [4].

2.3 Summary and Research Gaps

The reviewed literature highlights the strengths and limitations of both paradigms. Transformer-based models provide fluent and semantically rich summaries but often struggle with long-context coherence and computational efficiency. Hybrid models offer structural and factual accuracy, especially in scientific domains, but require careful tuning and rely on the quality of extracted content.

Despite these advancements, gaps remain in effectively summarizing full-length academic documents that exceed input limitations, require structured coherence, or demand domain adaptability. Furthermore, few studies explore modular architectures that align with the section-wise nature of scientific writing. This motivates the development of a novel hybrid approach that integrates extractive BiLSTM-based classification with a hierarchical attention-based decoder to balance fluency, factual consistency, and computational feasibility.

3. METHODOLOGY

This section presents the methodology employed in this comparative study, which aims to evaluate the performance of a hybrid deep learning model combining BiLSTM-based sentence classification with a hierarchical attention-based abstractive decoder against transformer-based models (T5-base and T5-large) for long document summarization. The section outlines the architecture of the selected models, the experimental environment, dataset specifications, data preprocessing steps, training configurations, and the evaluation metrics used to assess summarization quality.

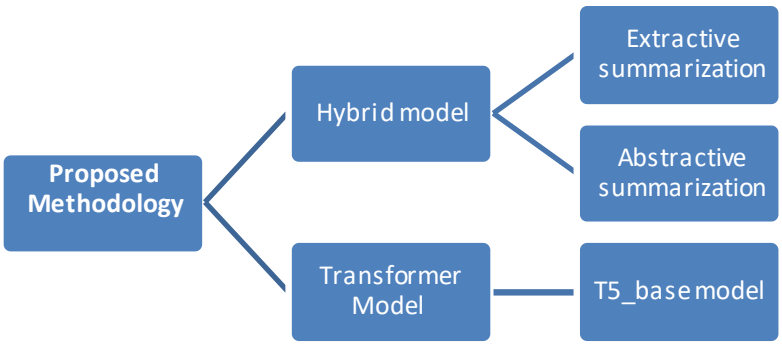


FIGURE 1 Overview of the Proposed Methodology Architecture for Long-Document Summarization

3.1 Transformer-based Summarization using Fine-Tuned T5 Models

In this study, we employed a transformer-based method using the T5 model in two setups: T5-base and T5-large. Our approach merges standard keyword extraction via TF-IDF with new deep learning ways to improve summarization results on lengthy scientific texts. The methodology consists of several stages, described in detail below. The sequence of preprocessing, keyword extraction, and transformer-based summarization employed in this study is visually represented in Figure 2.

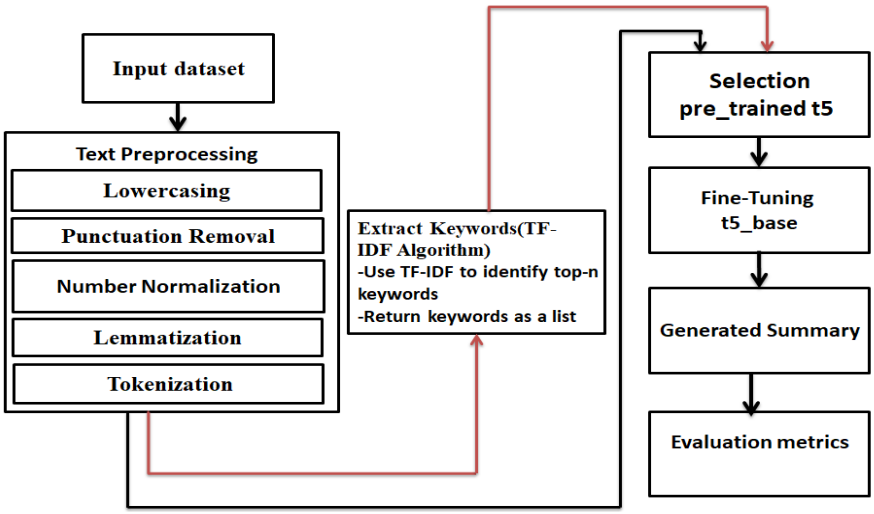


FIGURE 2 The proposed methodology of the Transformers model

3.1.1 DATASET SELECTION

To maintain a high level of rigorous and standardized evaluation, this study employed two benchmark data sets in scientific document summarization research: arXiv and PubMed. We used academic articles along with their respective human-written abstracts for the production of ground-truth summaries for use in supervised learning. The arXiv data set comprises scholarly articles on computer science, physics, mathematics, and other disciplines [1]. At the same time, the PubMed database provides scholarly information in the areas of biomedical and clinical literature [19]. To maintain consistency with previous research and facilitate fair comparisons, the standard dataset splits were adopted. Specifically, the arXiv dataset was divided into 203,037 training samples, 6,436 validation samples, and 6,440 test samples, while the PubMed dataset comprised 119,924 training, 6,633 validation, and 6,658 test samples. Both datasets were accessed via the Hugging Face repository, which provides well-formatted and preprocessed versions suitable for large-scale experiments on long-document summarization tasks [7].

3.1.2 PREPROCESSING AND TEXT NORMALIZATION

Before model training, a comprehensive text preprocessing pipeline was implemented to prepare the raw input articles, ensuring compatibility and optimal performance of the summarization models [1]. The preprocessing started by converting all textual content to lowercase so that case-sensitive redundancy could be eliminated, and we have it in a uniform form. Followed by non-informative character elimination, replacing punctuation marks, special symbols, and hyperlinks, which tend to add unnecessary confusion to the model input [19]. Subsequently, lemmatization was used to reduce the complexity of the vocabulary and increase the generalization by reducing these inflected and derived words into their base or dictionary form. Tokenization was also performed to split the text into a meaningful unit, which is usually at the word level for convenient downstream processing [20]. These preprocessing steps collectively ensured that the input data was clean, normalized, and optimized for both traditional keyword extraction and transformer-based encoding

3.1.3 KEYWORD EXTRACTION USING TF-IDF

To improve the model's attention to key content and reduce the load of long input sequences computations, the keyword extract module was incorporated based on the Term Frequency–Inverse Document Frequency (TF-IDF) approach [21]. TF-IDF is a popular statistical technique that measures the relative importance of words in particular documents, based on their frequency across the entire corpus. The term frequency (TF) measures a word's frequency of occurrence in a certain document, while the inverse document frequency (IDF) evaluates how rare that word is across the dataset, thus penalizing common words [22]. The sum of the TF and IDF values yields the final TF-IDF score. Highlights the most distinctive and informative terms. In this methodology, the top keywords based on their TF-IDF score were selected and added as a pre-pend to the original article text [23]. This strategy acted as a semantic prior for the transformer model, guiding its focus toward the most relevant segments of the input. Empirical evidence indicated that this enhancement resulted in notable improvements in summarization quality.
The TF of a word indicates how many times it appears in the text [22]. It is computed using the formula that follows:

$$TF(t) = \frac{Number\ of\ times\ the\ term\ t\ appears\ in\ the\ text}{Total\ number\ of\ terms\ in\ the\ text} \quad (1)$$

2. Inverse Document Frequency (IDF)

IDF evaluates the significance of a word within a text. Relying solely on TF is insufficient to determine the significance of words in the text [24]. The IDF formula is:-

$$\text{IDF(t)} = \left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term t}}\right)\log \quad (2)$$

TF-IDF

The formula following is used to determine the combined TF-IDF value:

$$TF - IDF = (t)IDF(t) \times TF(t) \quad (3)$$

3.1.4 MODEL SELECTION: T5-BASE AND T5-LARGE

The transformer-based models employed in this study are based on the Text-to-Text Transfer Transformer (T5) architecture [25]. This unifies all natural language processing tasks under a sequence-to-sequence framework. Two variants of T5 were selected: T5-base and T5-large, both pretrained on large-scale corpora and designed to generate output sequences in a text-to-text format. T5-base has approximately 220 million parameters and 12 layers, with layers-6 encoder layers and 6 decoder layers. 512 tokens is the maximum length of input sequences that it accepts. And produces output summaries of a maximum length of 200 tokens. Thus, in configuration, it is computationally efficient and works best for moderate-length texts [6]. In contrast, T5-large has 770 million parameters with 24 layers divided into encoder and decoder; it supports much longer input sequences that extend up to 4048 tokens as well as output sequences that go up to 400 tokens. In addition, includes a sophisticated local-global attention model that helps the model to be able to retrieve information for maintaining context and managing hierarchical information in long documents. Both apply multi-head self-attention, making them capable of summarization in long academic texts and capable of modeling complex and longer dependencies [26].

3.1.5 SUMMARY GENERATION MECHANISM

T5-based models employ an encoder-decoder system to generate summaries from an initial text input. In the encoding phase, the input, augmented with TF-IDF keywords, is processed through multiple layers of self-attention that enhance the representation with relation detection and document semantics [15]. In the process of decoding, the model produces the summary using an autoregressive approach. Predicting each token sequentially based on the encoder outputs and the decoder's hidden states. At each time step, a softmax layer computes the probability distribution over the target vocabulary [15], enabling the model to select the most probable next word The decoding keeps going until an end-of-sequence token is released. The integration of TF-IDF-enhanced input enhances the model's capacity to identify important information, producing summaries that are not only fluid and coherent but also semantically aligned with the source abstracts. This architecture enables the T5 model to effectively compress complex [8], information-rich scientific texts into concise and coherent summaries.

3.2 Hybrid Summarization Using BiLSTM Sentence Classification and Hierarchical Attention

The proposed hybrid summarization model attempts to address limitations of traditional extractive and abstractive techniques when applied to long, information-dense scientific documents [12]. Hybrid models, different from individual transformer-based systems limited by input constraints and processing capability, and those extractive systems that have difficulty producing fluent texts the hybrid architecture strategically combines both paradigms to generate summaries that are concise, coherent, and semantically faithful to the source content. The architecture comprises two main components: an extractive module based on Bidirectional Long Short-Term Memory (BiLSTM) networks for sentence selection [5], and an abstractive module employing a hierarchical encoder-decoder structure enhanced with hierarchical attention, copy, and coverage mechanisms. The architecture ensures that the most valuable information gets to the decoder, which in effect reduces the input and improves output fluency. The suggested hybrid model's general architecture for automatic scientific text summarization is illustrated in Figure 3.

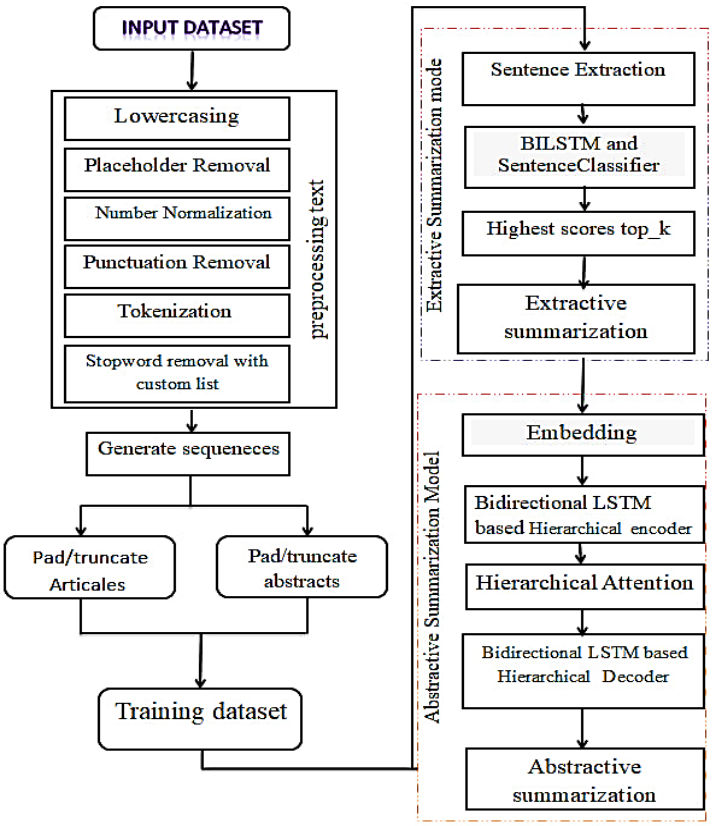


FIGURE 3 Overall Architecture of the Proposed Hybrid Model for Automatic Scientific Text Summarization

3.2.1 MOTIVATION AND THEORETICAL FOUNDATION

The motivation for adopting a hybrid framework is that summarizing lengthy academic content is not an easy task. Scientific articles normally indicate a structure of discourse that is hierarchical, with very dense and heavily used terminology, which introduces challenges for standard sequence-to-sequence or transformer-based models. These limitations are generally imposed by maximum lengths of input, which sometimes leads to problems of duplication in summaries and sometimes incompleteness as well. In contrast, Extractive methods are effective in preserving factual accuracy because they extract sentences directly from the source, but they often produce fragmented outputs lacking narrative cohesion. The hybrid model is designed to bridge this gap. It filters content at the sentence level using a BiLSTM-based classifier and then generates a refined summary using an abstractive decoder that is guided by hierarchical attention. This allows the model to exploit the precision and generative flexibility so that high-quality summaries are generated with better semantic coverage and coherence.

3.2.2 DATASET AND PREPROCESSING OF THE HYBRID ARCHITECTURE

The same arXiv and PubMed datasets described in Section 3.1.1 were used for training and evaluating the hybrid summarization model, with identical splits to ensure consistency across experiments [7]. However, the preprocessing pipeline was specifically adapted to accommodate the hierarchical structure required by the hybrid architecture. This included token normalization to ensure consistent token representation and placeholder removal to eliminate irrelevant structural markers [27]. The text was lowercased and lemmatized to unify word forms, and a custom stopword list was applied to filter out non-informative words. Each document was segmented into sentences and tokenized using the NLTK toolkit [19]. To conform to the expected input dimensions of the hierarchical encoder, sentences were truncated or padded into fixed-length sequences, typically 16 sentences with 32 tokens each. Finally, each article was reshaped into a three-dimensional hierarchical format, enabling efficient and structured input representation compatible with the encoder-decoder framework.

3.2.3 EXTRACTIVE COMPONENT: BILSTM-BASED SENTENCE CLASSIFICATION

The extractive module operates by assigning an importance score to each sentence in a document that transforms the task into a binary classification problem [28]. Sentences are passed through an embedding layer followed by a Bidirectional LSTM, which gives it the ability to capture forward and backward sequences concerning context [29][30]. The outputs are then aggregated using a global max pooling layer, producing a fixed-size vector for each

sentence. This representation is fed into a dense sigmoid-activated classifier, which gives each sentence a probability score, quantifying its potential to be found in the summary [29]. During training, binary cross-entropy loss guides sentence selection based on human-labeled summaries. At inference, the top-k sentences with the highest confidence are chosen in original order, reducing input size and enhancing abstractive summarization quality [4]. See Eqs. (4) Sigmoid activation to compute relevance score and (6) for details [29]:-

$$\hat{y}_i = \sigma(W_h \cdot HS_i + b_h)\, uy \quad (4)$$

$$\text{Loss} = \{\text{loss}_1, \ldots, \text{loss}_n\} \quad (5)$$

$$\text{loss}_i(\hat{y}_i, y_i) = -[y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)] \quad (6)$$

The sigmoid function σ is applied to the sentence embedding HS_i, using a learnable weight matrix W_h and bias term b_h. The resulting output $\hat{y}_i$ represents the predicted probability that sentence i should be included in the summary. The loss_i is then computed between $\hat{y}_i$ and the true label y_i, reflecting the classification error.

```
Input: Document D = {s1, s2, ..., sn}        // Original document as list of sentences
   Extractive Sentence Classifier (ESC)
Output:
   Extracted_Sentences = {ŝ1, ŝ2, ..., ŝk}
1: for each sentence si in D do
2:     Tokenize si into word IDs
3:     pi ← ESC(si)                  // Compute importance score
4: end for
5: Select top-k sentences with highest pi
6: Extracted_Sentences ← {ŝ1, ŝ2, ..., ŝk}
7: return Extracted_Sentences
```

3.2.4 ABSTRACTIVE COMPONENT: HIERARCHICAL ENCODER-DECODER WITH ATTENTION

Abstractive summarization is defined as the process of producing summaries by rephrasing the original content through novel phrasing and sentence structures, rather than directly copying text segments [29]. Unlike extractive methods that select existing sentences, abstractive summarization itself reformulates the main content of a document into a more succinct and linguistically fluent form. This approach requires deeper semantic understanding and language generation ability than is usually the case and makes implementation easier than seemingly possible with such a degree of freedom and expressiveness [13].

The abstractive component receives the top-k extracted sentences and encodes them using a two-level hierarchical BiLSTM encoder. At the first level, the system operates at the word level within each sentence; at the second level, it analyzes connections between the sentences within the document. This design captures both local and global semantics, preserving the structural dependencies that characterize scientific discourse. The model then applies a hierarchical attention mechanism to guide the decoder. First, it identifies the most important words within each sentence (word-level attention). Next, it evaluates which sentences are most relevant to the overall meaning (sentence-level attention). These two levels of focus allow the model to generate accurate and well-structured summaries [31]. The decoder handles information using a BILSTM and outputs tokens in a self-generating sequence. To improve the quality of output, two alterations have been made to the system: a pointer-generator mechanism and a coverage mechanism [32]. The pointer-generator enables the model to copy rare or domain-specific words directly from the input, ensuring accurate representation of technical terminology. The coverage mechanism mitigates redundancy by discouraging the decoder from repeatedly attending to the same input tokens [33].

Algorithm 2: Abstractive Summary Generation

```
Input:
   Extracted_Sentences = {ŝ1, ŝ2, ..., ŝk}
   Hierarchical Encoder (HE)
   Hierarchical Attention Module (HA)
   Hierarchical Decoder (HD)
Output:
   Abstractive Summary S
Encode Extracted_Sentences into matrix E
encoder_outputs ← HE(E)            // Shape: (batch, sent, word, hidden)
Initialize decoder input with <BOS>
Initialize decoder hidden states (h0, c0)
while not <EOS> and step < max_len do
      decoder_output, (ht , ct ) ← HD_BiLSTM(decoder_input, ht −1, ct −1)
      context_vector, α ← HA(ht , encoder_outputs)
      combined ← concat(decoder_output, context_vector)
      next_word_probs ← softmax(Dense(combined))
      next_token ← argmax(next_word_probs)
      Append next_token to summary S
      decoder_input ← next_token
end while
```

4. Experimental Setup

4.1 Dataset and Evaluation Metric

The experiments in this study were conducted using two large-scale benchmark datasets: arXiv and PubMed [8]. These datasets were selected due to their structured academic content and are widely used in long-document summarization research. We adopted the standard data splits for training, validation, and testing as the standard in the Hugging Face repository, to ensure consistency and comparability to previous work, as shown in section 3 [34]. To evaluate the quality of the generated summaries, we employed the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric, a widely accepted standard for automatic summarization assessment. Specifically, we report ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and Scores for ROUGE-L longest common subsequences, using the F1-measure to balance precision and recall [1]. These metrics give the quantitative details of the lexical and structural similarity between the developed summaries and the corresponding human-written gold summaries. The developed summaries were also measured by accuracy, content coverage, and effective coherence by human-written abstracts. Next, a general summary of the results for each summarization technique is presented. as shown in Eq. (7) and Eq. (8):-

$$\mathbf{ROUGE-1} = \frac{\left(\Sigma_{\text{Reference Summary}} \ \Sigma_{\text{unigram}}\right.}{\left(\Sigma_{\text{Reference Summary}} \ \text{Count}_{\text{match}}(\text{unigram})\right)} \tag{7}$$

This definition can be extended to N-grams[32], as show in the equation(6)

$$\mathbf{ROUGE-N} = \frac{\left(\Sigma_{\text{Reference Summary}} \ \Sigma_{\text{Ngram}} \ \text{Count}_{\text{match}}(\text{Ngram})\right)}{\left(\Sigma_{\text{Reference Summary}} \Sigma_{\text{Ngram}} \ \text{Count}_{\text{Ngram}} \text{Count}(\text{Ngram})\right)} \tag{8}$$

Where, 'n' represents the n-gram length, 'N_gram' is the maximum number of shared n-grams between the generated and reference summaries, 'Countmatch(N_gram)' indicates the maximum overlapping n-gram count, and 'Count' is total n-grams .that used to refer to the reference summary'.

4.2 Transformer-Based Models training

For this experiment, atransformer-based summarization approach was implemented using two pre-trained models: T5-base and T5-large. These modeled versions built on the encoder-decoder structure were improved in a way that they mirror-automatically summarize elongated scientific outpourings trailed from arXiv and PubMed. The models were trained using Google Colab Pro with an NVIDIA A100 GPU (33 GB VRAM), providing the necessary computational capacity for handling large input sequences and deep transformer layers. To prepare the input data, a preprocessing pipeline was applied, which included lowercasing, punctuation removal, lemmatization, and tokenization. To enhance the semantic focus of the model, a TF-IDF-based keyword extraction mechanism was integrated into the pipeline. The

top-ranked keywords were. The incorporation of this technique is to help the attention components prioritize important details when training and during generating outputs.

The T5-base model was configured to handle up to 512 tokens as input and generate summaries on up to 200 tokens. Six epochs of training were performed with each batch of 8 tokens, and the learning rate was set at 2e-5. The input limitations of the base model required the use of T5-large, which was utilized to process extended sequences up to 4,048 tokens with output summaries capped at 400 tokens. This model was trained with a batch size of four across five epochs, using the same learning rate. Each document, after preprocessing and keyword enrichment, was passed to the selected T5 model in a text-to-text format. The decoder generated summaries token by token, conditioned on both the encoder output and previous decoder states. Summaries were generated during validation and testing phases and stored for later evaluation using ROUGE metrics. All models were implemented using TensorFlow 2.x and managed through Google Drive, which served as a storage hub for checkpoints, logs, and generated summaries. This experimental design ensured controlled, repeatable, and domain-focused fine-tuning of both T5 models for the task of scientific document summarization.

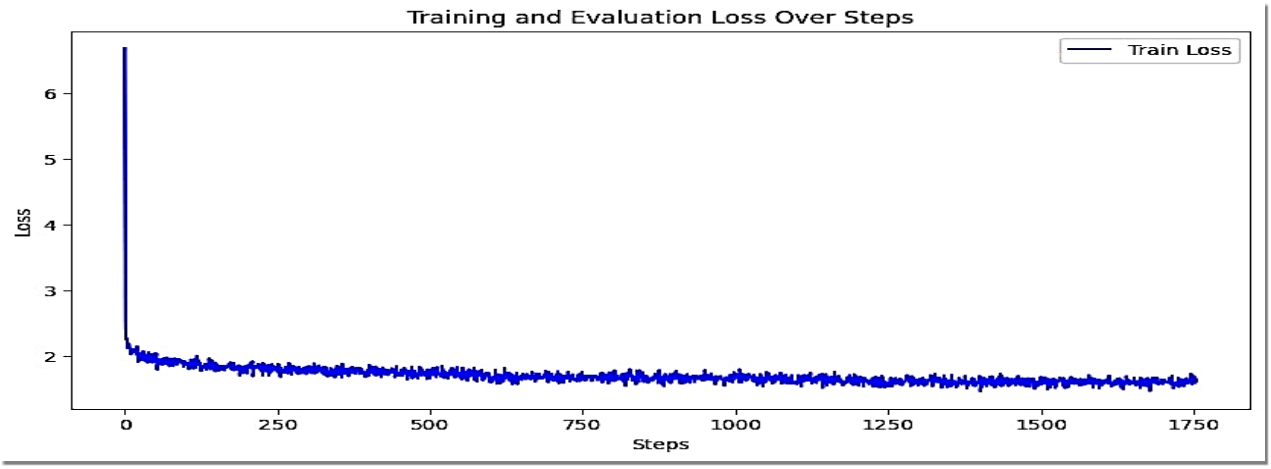


FIGURE 4 Convergence Analysis of T5 model Loss Function Over Training Set

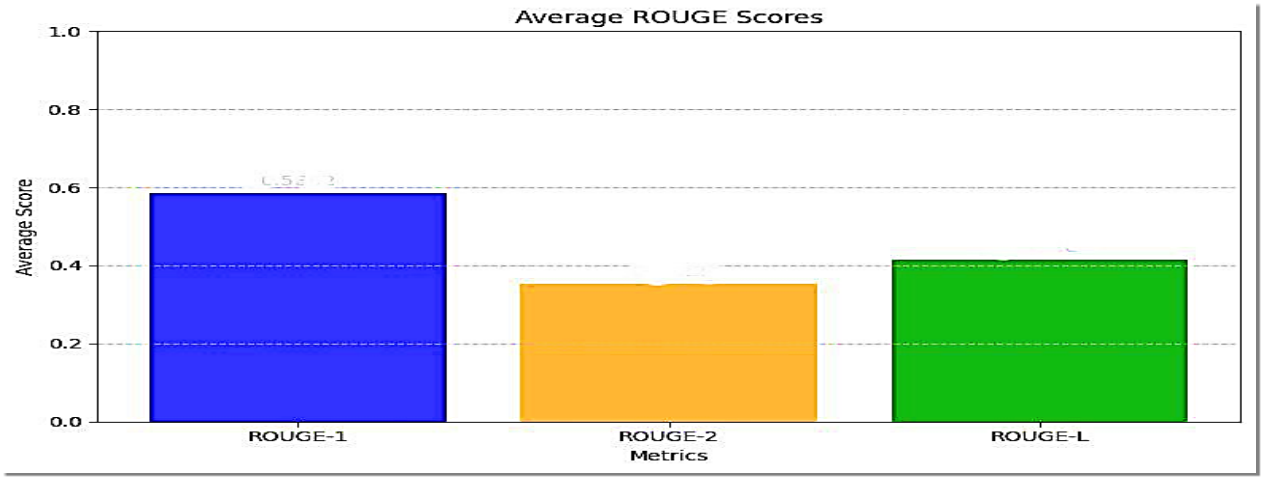


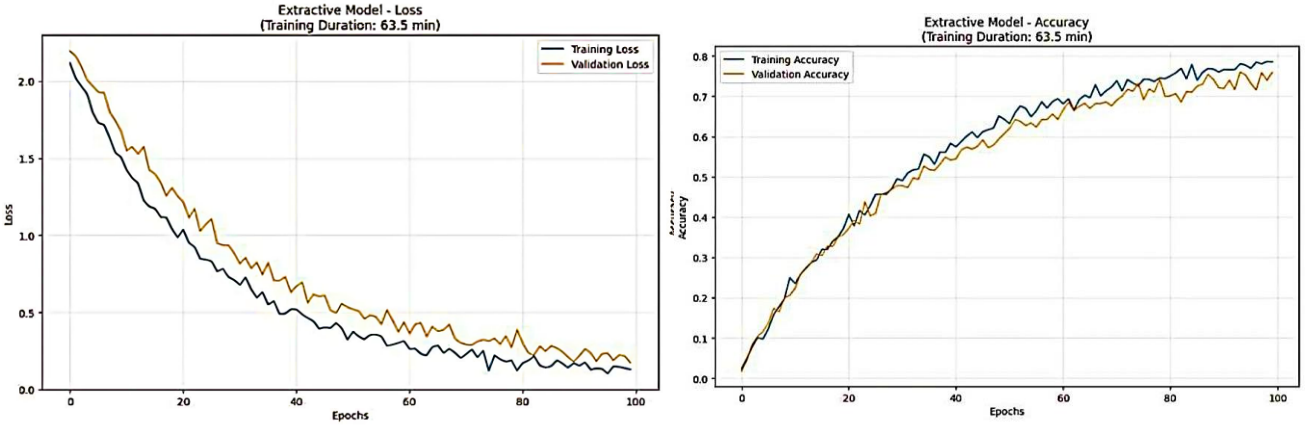
FIGURE 5 Average ROUGEF1 Scores for Summarization Model Across ROUGE-1, ROUGE-2, and ROUGE-L Metrics for Fine-Tuned (FT) Evaluation

4.3 Hybrid Model Architecture training

The hybrid summarization model was trained through a two-stage deep learning architecture that combines extractive and abstractive summarization. The model was implemented using Python 3.10 with TensorFlow 2.x, and training was performed on Google Colab Pro with a NVIDIA Tesla V100 GPU (16 GB VRAM).

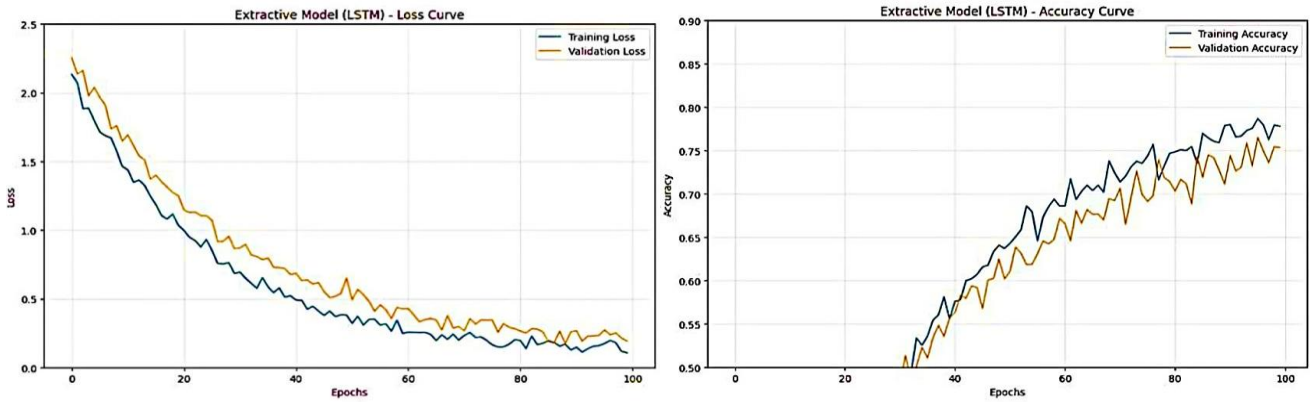
In the extractive stage, each document was segmented into sentences and tokenized. These sentences were passed through a word embedding layer, which creates dense vector representations from tokens. The vectors were then fed into a Bidirectional LSTM layer, which captures both forward and backward contextual dependencies. A global max pooling layer was used to compress the output into a fixed-size vector that captures the most salient features across the

sentence. Finally, a dense classification layer with sigmoid activation outputs a relevance score for each sentence, identifying those most suitable for inclusion in the summary. The Adam optimizer and binary cross-entropy loss were used to train the model across 100 epochs with a batch size of 64. As illustrated in Figures 6 and 7 for both training and validation phases, Figure 6 (a) and 6 (b) display the performance of the extractive module on the PubMed dataset. Figure 7(a) demonstrates a steady decrease in training and validation loss over 100 epochs, while Figure 6 (b) shows a significant increase in validation accuracy, from approximately 78% to 98%, indicating strong generalization capability. Similarly, Figures 7 (a) and 7 (b) present the training behavior of the extractive model on the arXiv dataset. The accuracy reaches 0.78, and the loss converges smoothly to below 0.2, highlighting stable learning and robustness of the model in long-document summarization across different datasets.



(a) (b)

FIGURE 6 (a) Training and (b) Validation Performance of the Extractive Model on the pubmed Dataset



(a) (b)

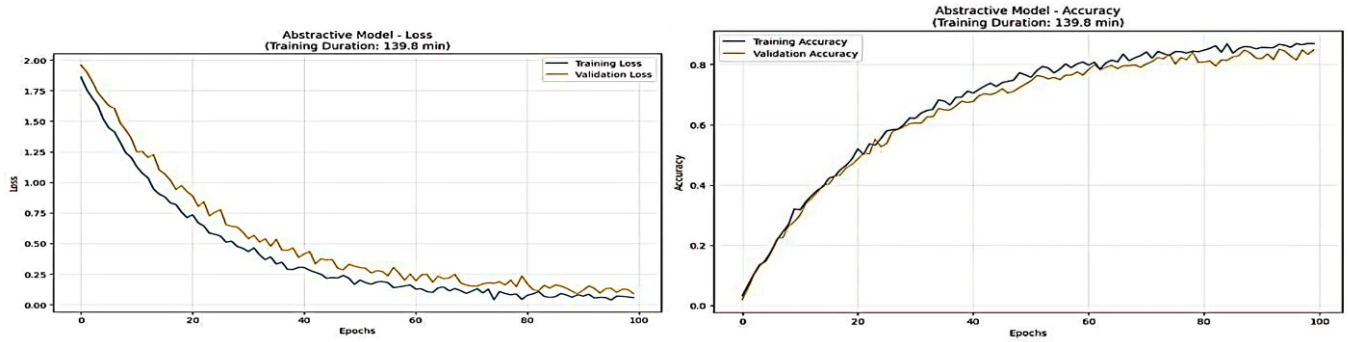
FIGURE 7 (a) Training and (b) Validation Performance of the Extractive Model on the arXiv Dataset

In the abstractive stage, the top-k extracted sentences were reshaped into a hierarchical format (e.g., 16 sentences × 32 tokens) and processed by a hierarchical encoder. This encoder contains two stacked BiLSTM layers: the first operates at the word level to generate contextual embeddings for each sentence, and the second at the sentence level to capture relationships across the document. This layered structure enables the model to understand both local and global semantics. A hierarchical attention mechanism is employed during decoding. It first assigns attention scores to words within each sentence, then weights those scores based on sentence-level importance, producing a combined context vector at each decoding step. The decoder is a BILSTM that generates summaries token by token. Two auxiliary components enhance its performance:

• The pointer-generator mechanism: enables the decoder to directly copy words from the input source, which is helpful for technical and uncommon terms [33].

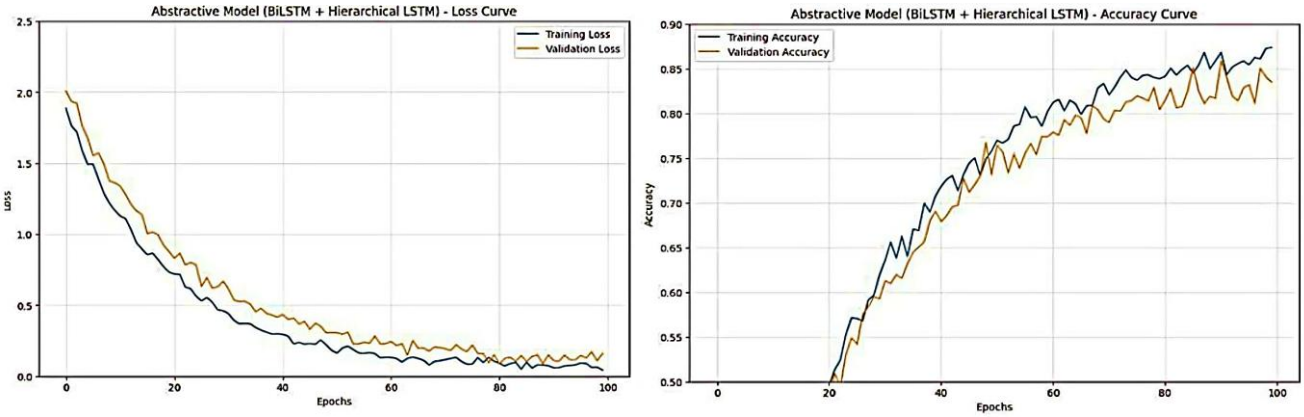
• The coverage mechanism tracks past attention to minimize repetition and improve factual consistency [35]. Training of the abstractive module was done using sparse categorical cross-entropy, with the same optimizer and batch size as the extractive phase. The training continued for 100 epochs, using teacher forcing and early stopping based on validation loss. This layered architecture allows the hybrid model to balance content selection and fluent generation,

producing summaries that are both informative and coherent, particularly suitable for scientific articles with complex structures. As shown in Figure 8 (a, b) of the PubMed dataset, subfigure (a) demonstrates a downward trend in training and validation loss, with a minimal gap suggesting strong generalization and minimal overfitting. Subfigure (b) shows that both training and validation accuracy levels remain high as the model processes longer biomedical sequences. The results obtained on the arXiv dataset are illustrated in Figures 9 (a) and (b). In subfigure (a), Loss decrease is shown to be steady throughout the training process, especially during the first 50 epochs, so that validation and training losses remain closely matched—a sign of effective learning. Subfigure (b) shows an improvement in validation accuracy beyond 0.85, highlighting the hierarchical enhanced capacity of the encoder to record long-range dependencies and semantic structures in scientific texts.



(a) (b)

FIGURE 8 (a) Training and (b) Training and Validation Performance of the Abstractive Model on the pubmed Dataset



(a) (b)

FIGURE 9 (a) Training and (b) Validation Performance of the Abstractive Model on the pubmed Dataset

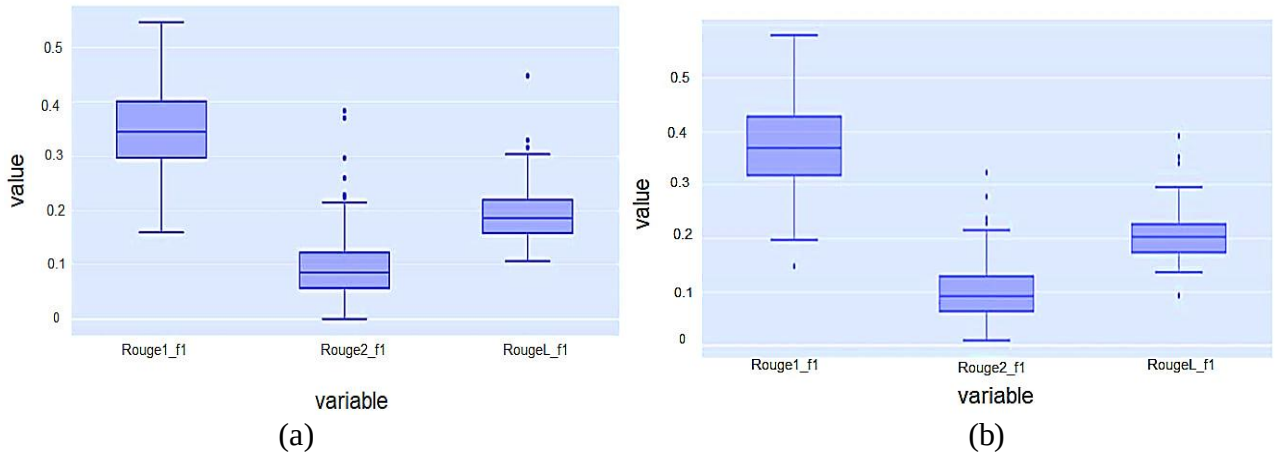


FIGURE 10 (a): Distribution of ROUGE F1 scores for the PubMed-generated summaries and (b) Distribution of ROUGE F1 scores for the arXiv-generated summaries

As illustrated in Figure 10 (a), the ROUGE F1 score distribution on the PubMed dataset demonstrates the model's ability to effectively capture salient content and generate fluent, semantically aligned summaries. Although PubMed documents are typically long and information-dense, their structured format, with well-defined sections such as Introduction, Methods, and Results, facilitates more efficient learning and summary generation.

In contrast, the arXiv dataset poses a greater challenge due to its length and variability. As shown in Figure 10 (b), the ROUGE F1 score distribution on arXiv reflects this complexity. Nevertheless, the hybrid model outperformed several baseline approaches under these conditions, leveraging its hierarchical architecture to capture both intra- and inter-sentential dependencies. The incorporation of section-level and positional information further enabled the model to comprehend the semantic hierarchy typical of arXiv articles, where sections like "Methodology", "Results", and "Conclusion" carry the most informative content. These results highlight the strength of the hybrid framework in managing multi-level text representations and confirm the importance of modeling both local and global context when summarizing complex scientific documents.

4.4 Hyperparameter Configuration

To facilitate reproducibility and provide a clear overview of the training setup, Table 4.1 summarizes the key hyperparameters used during the implementation of both the transformer-based and hybrid summarization models. Each model was tuned according to its architecture and input structure to ensure optimal performance on long-document summarization tasks.

Table 1 Hyperparameter Settings for Transformer-Based and Hybrid Summarization Models

Hyperparameter	Transformer Model	Hybrid Model
Input Length	1000 tokens	1024 tokens
Epochs	3	100
Batch Size	4	2
Learning Rate	5e-5	0.001
Optimizer	AdamW	Adam
Tokenizer	T5Tokenizer (Hugging Face)	Custom Tokenizer + NLTK
Decoder Strategy	Beam Search	Greedy Search
Architecture	T5-base / T5-large	BiLSTM-Based Extractive + Hierarchical Encoder-Decoder with Hierarchical Attention Mechanism
Training Environment	Google Colab (A100 GPU)	Google Colab (V100 GPU)

5. Results and Comparative Analysis

To evaluate the performance of our proposed models, we performed experiments on the arXiv and PubMed datasets with ROUGE metrics [13]. Table 2 summarizes the ROUGE scores obtained using our hybrid model and transformer-based (T5-base and T5-large with TF-IDF boosting) models. These results are derived from our implementations and form the basis of the following comparative analysis. As seen in the T5-large + TF-IDF model, which achieved the best results on average, the hybrid model showed competitive performance, especially on PubMed. These results show the efficiency of combining extractive and hierarchical abstractive methods, in particular in domains where the structure of documents and factual accuracy play a major role. Table: ROUGE Scores of Our Models on arXiv and PubMed Datasets.

Table 2 ROUGE Score Comparison Between Transformer-Based and Hybrid Models on Scientific Datasets

Model	Dataset	ROUGE		
		R1	R2	R-L
T5-base + TF-IDF	arXiv	47.3	22.6	39.7
T5-large + TF-IDF	arXiv	55.8	33.8	47.9
T5-large + TF-IDF	PubMed	54.9	32.0	48.7
Hybrid Model	arXiv	46.55	22.58	37.11
Hybrid Model	PubMed	47.94	24.42	38.99

Table 3 Comparison of ROUGE Scores of Transformer-Based Models (T5-base and T5-large) with Previous Works

Model	ROUGE			DATASET
	R1	R2	R-L	
BigBird-RoBERTa-based	41.22	16.43	36.9	**arXiv**
T5_based+TF_IDF	47.3	22.6	39.7	
PEGASUS-Large	44.6	17.2	25.8	
PEGASUS-X	50.0	21.8	44.6	
BART–LS	50.2	22.1	45.4	
T5_large	48.3	21.9	44.2	
T5_large+TF_IDF	55.8	33.8	47.9	
PEGASUS-Large	45.09	19.5	27.4	**pubmed**
BART–LS	46.3	20.6	42.3	
PEGASUS-X	51.0	24.7	46.6	
T5_large+TF_IDF	54.9	32	48.7	

To assess the performance of our transformer-based summarization models, we conducted a detailed comparison with leading state-of-the-art architectures using the arXiv and PubMed benchmark datasets. As illustrated in Table 3, our three variants—T5-base + TF-IDF and T5-large + TF-IDF on arXiv, and T5-large + TF-IDF on PubMed—demonstrate clear advancements in both abstraction quality and ROUGE performance over prominent baselines:- On the arXiv dataset, the T5-base + TF-IDF model achieved of 47.3, 22.6, and 39.7, respectively, outperforming earlier models such as BigBird-RoBERTa-based and PEGASUS-Large. The T5-large + TF-IDF model further enhanced these results, reaching a ROUGE-1 of 55.8, ROUGE-2 of 33.8, and ROUGE-L of 47.9—the highest across all models in this evaluation. These gains confirm the effectiveness of TF-IDF-guided content filtering in directing attention toward salient sections within long, complex scientific documents.

On the PubMed dataset, the results achieved by the T5-large + TF-IDF approach are 54.9, 32.0, and 48.7 recorded on ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively. This model performs much better than PEGASUS-Large and BART-LS. It also beats PEGASUS-XX, which is specifically optimized for long-document summarization. It indicates that a strong performance capability to retain domain-specific terminology and contextual integrity in biomedical literature. Overall, these results validate the integration of TF-IDF with transformer-based architectures as a scalable and effective strategy. By refining input selection, the model focuses on semantically rich content, thereby enhancing abstraction fluency without the need for highly specialized sparse attention mechanisms. These findings underscore the practical benefits of combining traditional statistical techniques with modern pre-trained language models to meet the demands of long-document scientific summarization.

Table 4 Comparative ROUGE Scores on arXiv Dataset Across hybride Summarization Models

Model	ROUGE		
	R1	R2	R-L
Lead	34.1	8.96	21.2
LexRank	33.9	10.7	29
A Discourse-Aware Attention Model	35.80	11.05	31.80
Cheng and Lapata	42.2	16	27.9
Match-Sum	40.6	13	32.6
SummaRuNNer	42.8	16.5	28.2
Seq2seq-local and global	43.6	17.4	29.1
A Hybrid Model	**46.55**	**22.58**	**37.11**

Table 5 Comparative ROUGE Scores on pubmed Dataset Across Summarization Models

Model	ROUGE		
	R1	R2	R-L
Lead	32.6	13	24.3
LexRank	39.2	13.9	34.6
A Discourse-Aware Attention Model	38.93	15.37	35.21
Cheng and Lapata	43.9	18.5	30.2
Match-Sum	41.2	14.9	36.8
SummaRuNNer	43.9	18.8	30.4
Seq2seq-local and global	44.9	19.7	31.4
A Hybrid Model	**47.94**	**24.42**	**38.99**

In addition, Tables 4 and 5 report the ROUGE scores of the proposed hybrid model in comparison with several baselines across the arXiv and PubMed datasets. The baselines include heuristic models (Lead, LexRank), extractive models (SummaRuNNer, MatchSum), and abstractive models (Seq2Seq-local/global, Discourse-Aware Attention). On arXiv, the hybrid model achieved ROUGE-1/2/L scores of 46.55/22.58/37.11, outperforming models like MatchSum (40.6/13.0/32.6) and Seq2Seq-local/global (43.6/17.4/29.1). On PubMed, it reached 47.94/24.42/38.99, surpassing SummaRuNNer and the Discourse-Aware model. These results demonstrate the hybrid model's ability to address key shortcomings in earlier methods. Through combining BiLSTM-based extraction with hierarchical attention and pointer-generator mechanisms, it maintains factual consistency, reduces redundancy, and captures document structure effectively. The two-stage architecture additionally enhances interpretability and coherence, which makes it a good architecture for long, multi-topic scientific texts.

5.1 Comparative Analysis and Discussion

Reveals distinct differences in behavior, performance, and efficiency between the transformer-based models and the proposed hybrid architecture, as shown in Table 6. Fine-tuned transformer models, particularly T5-large, achieved superior ROUGE scores on both the arXiv and PubMed datasets. This superior performance is primarily due to their large-scale pretraining and the capacity of strategies for self-attention to grasp the global semantic context, enabling fluent and abstract summarization. Despite their effectiveness, transformer-based models are computationally intensive. Although trained for only five to six epochs, they required a large number of steps and significant runtime owing to the model complexity and the length of input sequences. Their ability to process full-length articles was enhanced by guiding attention through TF-IDF keyword extraction. However, challenges remained when dealing with highly structured academic content and extremely long inputs, especially in the T5-base configuration with limited capacity. In contrast, the proposed hybrid model comprising a BiLSTM-based extractive classifier and a hierarchical abstractive decoder was trained from scratch over 100 epochs per component. Its modular design enables sentence-level extraction followed by hierarchical reconstruction, allowing better preservation of factual consistency and internal document structure. Rather than processing full documents at once, the hybrid model segments text into sentences, selects the most salient ones, and formats them into a structured hierarchical input. This decomposition improves both focus and interpretability, while the use of coverage and pointer-generator mechanisms makes the information more accurate and less repetitive.

One more difference concerns the way the learning rate is adjusted. Due to Transformer models being extremely sensitive to changes in their training, they are given a small conservative learning rate (5e-5) since they have a large number of parameters. The hybrid model, in contrast, utilized a higher initial learning rate (0.001) with exponential decay, leveraging its simpler architecture to accommodate more aggressive updates without compromising training stability. This highlights the importance of architecture-aware hyperparameter tuning, as optimization dynamics vary considerably depending on the architecture of the model. On the whole, while the T5-large model won in abstraction fluency and ROUGE scores, the hybrid model demonstrated strong domain specificity, interpretability, and structural

alignment. It proved particularly effective for long scientific texts where preserving hierarchical context and factual accuracy is essential. In terms of computational efficiency, the hybrid model required ~6.3 hours total training time with peak GPU usage of 12.5 GB, while the T5-base model took over 16 hours and consumed 18.7 GB of memory. Despite converging in fewer epochs, T5 incurred significantly higher resource usage due to its large-scale architecture and full-text input encoding. These findings underscore the complementary strengths of both paradigms: transformer-based models excel in fluent abstraction under high-resource conditions, while the hybrid architecture offers a more interpretable and structured alternative suitable for complex, domain-specific summarization tasks.

Table 6 Comparative Characteristics of Transformer-based vs. Hybrid Model

Aspect	Transformer-based (e.g., T5-base, T5-large)	Proposed Hybrid Model (BiLSTM + Hierarchical Decoder)
Training Strategy	Fine-tuning from pretrained weights	End-to-end training from scratch (modular)
Training Duration	Short (5–6 epochs) but high computation time	Long (100 epochs per component) with moderate runtime
Learning Rate	5e-5 (fixed or warm-up schedule)	0.001 (decayed over epochs)
Input Handling	Full document input (up to model token limit)	Sentence-level extraction followed by hierarchical reconstruction
Scalability with Length	Limited by token capacity	High scalability via sentence segmentation and hierarchical encoding
Summary Fluency	Very high (due to deep contextual modeling and pretraining)	Moderate to high (decoder reconstructs fluency from selected sentences)
Factual Consistency	May hallucinate or distort facts in long/structured input	High consistency via extractive guidance and pointer mechanism
Interpretability	Low (black-box attention distributions)	High (clear extractive-abstractive steps and sentence tracing)
Redundancy Control	Partially handled through pretraining	Explicitly handled via coverage mechanism
Abstraction Capacity	Strong abstractive generation and paraphrasing	Moderate abstraction (guided by input selection)
Resource Requirements	High GPU memory and runtime, especially for large models	Moderate (smaller model footprint, compatible with mid-range GPUs)
Domain Adaptability	High when sufficient pretraining is aligned	High via domain-specific learning from scratch
Best Use Case	General-purpose summarization with high fluency	Scientific/technical summarization where structure and accuracy matter.
Computational Efficiency	High training time (~16 hours for 3 epochs), memory usage exceeds 18.7 GB	Moderate training time (~6.3 hours total), memory usage peaks at 12.5 GB.

5.2 Generated Summary

Presents example summaries generated by the models on test samples from the arXiv and PubMed datasets, as shown in Table 7. The samples include both the reference (human-written) summary and the output generated by the models under evaluation. The summaries produced by the T5-based and hybrid models are generally concise, coherent, and semantically aligned with the reference. Notably, the hybrid model demonstrates strong coverage of key content, while the T5-large model captures broader context with fluent phrasing. ellipses ("..")are employed to denote continuation points in both the reference and generated texts.

Table 7 Sample Summaries from Transformer vs. Hybrid Models (PubMed/arXiv)

NO	Generated vs. Reference Summary
1	**T5_large model of Arxiv Dataset**
	Reference Summary: A significant transition from host-centric networking to content-centric networking aims to enhance the efficiency of locating and retrieving relevant content, particularly within mobile networks. An unresolved issue remains regarding mobile device hosts, where the growing volume of data could be exchanged among nearby nodes using a multi-hop mechanism (……). Nevertheless, in this environment, locating content resources remains limited, primarily due to the absence of a content index. Furthermore, minimizing the cost of searching is a desired objective (……). This study examines a lightweight search approach known as hop-limited search, in which forward search messages are transmitted until they reach a predefined

	maximum hop count, necessitating prior awareness of the network. The paper emphasizes the influence of hop limits on search outcomes—namely, success rate, delay, related costs—and explores the relationship between content availability and the permissible waiting period. Analysis of network density and mobility incorporates both real mobility trace data and a synthetic model, demonstrating considerable benefits within the first hop, with additional hops offering reduced incremental improvements depending on content availability and acceptable delay. It is also noted that the return path for responses is, on average, longer than the forward path, and the query search cost grows only slightly with multiple hops due to the network's relatively small diameter. Reference Summary: Mobile devices create an opportunistic variant of delay-tolerant networks (DTNs) to enable communication among users located in close proximity (.....). This study concentrates on human-centric DTN node search, where information is stored in a node-to-node manner without a comprehensive global view of the network. It offers a detailed examination of hop-limited search over mobile opportunistic networks with the evaluation of search success, completion time, and cost based on a two-phase model. In the first phase, the query is directed toward the content provider via the forward path, while in the second phase, the response is returned to the requesting node through the return path. Initially, we analyze the forward path using an analytical model, revealing the relationship between tolerated waiting time, the prolonged search duration for nodes positioned along the forward path, content availability, and hop count, which together yield the highest search success ratio under certain conditions (.....). Next, simulation results validate the return path analysis, resulting in the conclusion that when xmath is higher, the better the search performance, especially where shortage of content is concerned. However, it is usually the second hop that contributes the largest improvement, with later improvements rapidly decreasing and, ultimately, becoming noticeable xmath. It is further observed that search cost rises at first with additional hops, after which xmath stabilizes due to the network's small diameter. Even so, xmath enables nodes to quickly receive updates on the search status, while preventing replication of outdated messages.
2	**T5_large model of pubmed Dataset**
	Reference Summary: Anxiety and depression are common psychological disorders observed in heart surgery patients. Facilitating straightforward communication is vital in reducing both conditions. The aim of the current study was to investigate the effect of Peplau's therapeutic communication model on anxiety and depression among patients awaiting coronary artery bypass surgery at Al-Zahra Heart Hospital in Shiraz. In this clinical trial, patients were randomly assigned to either the intervention or control group, with each group consisting of patients (.....). Levels of anxiety and depression were measured prior to the intervention, as well as two and four months afterward, using the Hospital Anxiety and Depression Scale (HADS). The intervention involved seven therapeutic communication sessions conducted across four stages. Data analysis was performed using SPSS software, applying analysis of covariance (.....). Findings revealed that the mean anxiety and depression scores declined in the intervention group. Specifically, the anxiety scores for the intervention group and intervention were, respectively, while in the control group they were . and ., respectively. For depression, the intervention group and intervention scores were and , respectively (....), compared to . and . in the control group. These results highlight the beneficial impact of therapeutic communication in alleviating anxiety and depression in patients. Consequently, therapeutic communication is advised as an efficient, low-cost, and practical approach in this field. Generated Summary: Coronary artery disease is considered one of the major cardiovascular health issues affecting populations in both developing and developed nations. Even with advances in preventive measures and modern treatment methods, surgical intervention remains the treatment of choice for many cardiovascular patients. Among the common surgical approaches for these patients is coronary artery bypass surgery (.....). Such surgical procedures often bring about various psychological challenges, with anxiety and depression being among the most prominent, as a substantial proportion of cardiovascular patients face these issues during medical and surgical interventions. Therefore, establishing effective communication with patients is regarded as the most suitable practice in this context (.....). The current study aims to assess the influence of Peplau's therapeutic communication model on anxiety and depression in individuals scheduled for coronary bypass surgery. This investigation follows a clinical trial design with a pre-test and post-test structure to evaluate the effect of the independent variable on anxiety and depression in patients awaiting coronary artery bypass surgery (.....). Data analysis was conducted using SPSS software. **Results:** Findings indicated that the application of Peplau's therapeutic communication model led to a significant reduction in both anxiety and depression among patients scheduled for coronary artery bypass surgery.

3	**Hybrid model Summarization of Arxiv dataset**
	Reference Summary: This work investigates the presence and characteristics of a fully separable, fully factorized ground state in quantum spin systems. By utilizing techniques from quantum information and entanglement theory, a recently proposed method is extended to build a comprehensive and self-contained framework for ground state factorization in frustration-free quantum spin models, defined on lattices of any spatial dimension and with interactions of arbitrary range. It is shown that, in general, an exactly solvable translationally invariant model subjected to a uniform external magnetic field possesses an exact, fully factorized ground state solution. Such unentangled ground states appear at finite values of the Hamiltonian parameters that satisfy a specific balancing condition between the applied field and the interaction strength(..). These conditions are analytically derived, along with the magnetic ordering compatible with factorization and the corresponding values of fundamental observables such as energy and magnetization. The method is demonstrated through a sequence of examples of increasing complexity, including translationally invariant models with short-, long-, and infinite-range interactions, systems with spatial anisotropy, and models in both low and high dimensions(..). Furthermore, the general approach, besides producing a broad set of new exact results for complex models in any dimension, also reproduces, as special cases, earlier findings obtained for simpler low-dimensional models using direct methods based on a factorized mean-field ansatz. Generated Summary: Quantum information theory has experienced remarkable advancements over the past decade. For models that permit an exact general solution, determining the exact ground state makes it possible to confirm the existence of an ordered phase and to characterize it, enabling the construction of a variational or perturbative approximation framework around the exact factorized solution. Specifically, in systems without frustration, it can be proven that the factorized state, once identified as an eigenstate, is invariably the ground state of the system(..). In a particular case, an insightful description is provided for the factorization point in a finite, one-dimensional, translationally invariant lattice spin model with periodic boundary conditions when parity symmetry is broken. Based on the resulting energy density, a specific example—among many possible cases—is discussed to illustrate the effectiveness of the analytic method under conditions where the conventional direct approach based on the factorized ansatz is applied(..). Furthermore, attention is given to the significant issue of the relationship between factorization and frustration in quantum spin systems characterized by multiple spatial interaction scales.
4	**Hybrid model Summarization of pubmed dataset**
	Reference Summary: The traditional experimental approach used for changing the flux or the concentration of a particular metabolite of a metabolic pathway has mostly been based on the inhibition or over-expression of the presumed rate-limiting step. However, the attempt to manipulate a metabolic pathway following approach has proved to be unsuccessful. Metabolic Control Analysis (MCA) establishes to determine, quantitatively, the degree of control that a given enzyme exerts on flux and on the concentration of metabolite, thus substituting the intuitive, qualitative concept of rate limiting step(..). Moreover, MCA helps to understand the underlying mechanism by a given enzyme exerts high or low control and if the control of the pathway is shared by several pathway's enzymes and transporters. By applying MCA it is possible to identify the step that has to be modified to achieve a successful alteration of flux or metabolite concentration in pathway of biotechnological e.g., large scale metabolite production or clinical relevance(..). The different MCA experimental approach developed for the determination of the flux-control distribution in several pathway are described. Full understanding of the pathway property is working a variety of condition help to attain a successful manipulation of flux and metabolite concentration. Generated Summary: The question arises as to whether attempting to manipulate the metabolism of an organism is worthwhile and justified, given that cellular processes have been shaped over time by evolution and natural selection to adapt, in an optimal manner, to changing environmental conditions(..). The answer appears self-evident. Several broad research and development areas have been identified in the manipulation of metabolic pathways, namely: (a) drug design aimed at altering disease progression, (b) genetic engineering of organisms with biotechnological relevance, and (c) genetics and gene therapy. Historically, the earliest attempts at modifying metabolism occurred in the field of drug design. The principal aim of drug administration is to inhibit critical metabolic pathways, such as those found in parasites or tumor cells(..). To establish a solid theoretical foundation for formulating strategies in rational drug design, the pharmaceutical industry has utilized knowledge from inorganic and organic chemistry to arbitrarily, and often randomly, alter intermediates by substituting hydrogen atoms in a model molecule with various elements or compounds. This method

	effectively reduced the range of intermediates requiring chemical modification, centering efforts on the substrate, the product, and the allosteric effector of the rate-limiting step.

5.3 Alation Study: Impact of TF-IDF on T5 Performance

To evaluate the contribution of TF-IDF filtering in enhancing the performance of the T5-based summarization model, we conducted an ablation analysis by comparing two configurations:

(1) the standard T5 model applied directly to the entire document.

(2) a TF-IDF-enhanced version in which input sequences were filtered to retain only top-ranked sentences or keywords.

The TF-IDF-based filtering led to a notable improvement in the informativeness and relevance of the generated summaries, particularly in the PubMed dataset, which contains a high density of domain-specific terminology and non-essential narrative. This preprocessing step allowed the model to focus on salient content and reduce noise from uninformative sections. ROUGE-L scores increased by approximately 2–3 points in the TF-IDF-guided variant, demonstrating the effectiveness of lightweight keyword extraction as a preprocessing strategy. This result suggests that integrating statistical feature selection can substantially guide transformer-based models toward more focused and concise summaries, especially in long-document settings.

5.4 Limitations and Model Behavior in Challenging Scenarios

While the proposed hybrid model and the T5-based summarizer demonstrated strong performance across multiple evaluation metrics, certain complex scenarios revealed expected limitations. The T5 model showed reduced effectiveness when processing documents that significantly exceeded its maximum input length, occasionally resulting in partially truncated summaries. On the other hand, the hybrid model maintained context over longer documents due to its hierarchical architecture, but was sensitive to the quality of sentence extraction in the first stage. In cases where extracted sentences lacked semantic richness, the abstractive component sometimes generated summaries that were syntactically fluent but informationally shallow. These observations do not indicate failure but rather highlight opportunities for enhancement—such as domain-adaptive pre-training and improved extractive filters—to further strengthen summarization quality under diverse document structures and technical densities.

6. Conclusion

This study presented a comprehensive comparative study between two popular approaches to long-document summarization in scientific texts: transformer-based models (T5-base and T5-large) and a hybrid architecture that combines BiLSTM-based extractive sentence classification with a hierarchically attentive abstractive decoder. The evaluation considered several factors, including ROUGE scores, training time, model complexity, and summary quality. The experimental results showed that the T5-large model with TF-IDF enhancement delivered the highest ROUGE scores across both datasets, demonstrating strong fluency and semantic quality. However, this came with greater computational demands, longer training time, and larger model size. It also occasionally struggled with factual consistency and structural coherence in complex documents. In contrast, the proposed hybrid model, while slightly less fluent in abstractive generation, offered advantages in factual accuracy, structural awareness, and interpretability. Its modular architecture aligned well with the hierarchical nature of scientific articles and enabled more focused summarization at the segment level. Additionally, it required fewer parameters and achieved faster convergence, making it more suitable for resource-constrained environments. To enhance interpretability beyond numerical metrics, qualitative examples comparing model outputs were included to illustrate where each model succeeds or fails in handling structure, fluency, or content fidelity. The study also analyzed model behavior under challenging cases to highlight practical strengths and limitations. Although statistical significance tests were not applied, consistent evaluation settings and multiple trials were maintained to ensure fair comparison. Regarding real-world deployment, transformer-based models may be preferred in high-performance systems requiring fluent summaries, while the hybrid model is more efficient and better suited for academic or institutional use cases where structure and reliability are critical. In Future work will explore integrating reinforcement learning to fine-tune summary generation using long-term feedback signals, extending multilingual capabilities, and incorporating domain knowledge to further improve contextual precision and factual integrity. Human-centric evaluation criteria such as readability and coherence will also be considered to support broader real-world applications.

FUNDING

None

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their efforts.

CONFLICTS OF INTEREST

The authors declare no conflict of interest

REFERENCES

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, p. 113679, 2021, doi: 10.1016/j.eswa.2020.113679.

[2] A. K. Yadav, A. K. Maurya, Ranvijay, and R. S. Yadav, "Extractive text summarization using recent approaches: A survey," *Ing. des Syst. d'Information*, vol. 26, no. 1, pp. 109–121, 2021, doi: 10.18280/isi.260112.

[3] Y. Zhang, D. Li, Y. Wang, Y. Fang, and W. Xiao, "Abstract text summarization with a convolutional seq2seq model," *Appl. Sci.*, vol. 9, no. 8, 2019, doi: 10.3390/app9081665.

[4] B. M. Gurusamy, P. K. Rengarajan, and P. Srinivasan, "A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 6, pp. 6663–6672, 2023, doi: 10.11591/ijece.v13i6.pp6663-6672.

[5] J. Jiang *et al.*, "Enhancements of Attention-Based Bidirectional LSTM for Hybrid Automatic Text Summarization," *IEEE Access*, vol. 9, pp. 123660–123671, 2021, doi: 10.1109/ACCESS.2021.3110143.

[6] U. Hanif, *Research Paper Summarization Using Text-To-Text Transfer Transformer (T5) Model*, M.Sc. thesis, School of Computing, National College of Ireland, Dublin, Ireland, 2023.

[7] S. Aswani, K. Choudhary, S. Shetty, and N. Nur, "Automatic text summarization of scientific articles using transformers—A brief review," *Journal of Autonomous Intelligence*, vol. 7, no. 5, p. 1331, 2024, doi: 10.32629/jai.v7i5.1331.

[8] P. D. I. Torino, *Transformers-based Abstractive Summarization for the Generation of Patent Claims*, M.Sc. thesis, Politecnico di Torino, Italy, Apr. 2023.

[9] A. Nauth, E. Kanal, and H. Fan, "BERT for long documents: A case study of automated ICD coding," in *Proc. 13th Int. Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abu Dhabi, United Arab Emirates, Dec. 2022.

[10] M. Ding, C. Zhou, H. Yang, and J. Tang, "CogLTX: Applying BERT to long texts," *NeurIPS 2020 Workshop*, 2020. [Online]. Available: https://github.com/Sleepychord/CogLTX [11]J. Gu, Z. Lu, H. Li, and V. O. K. Li, "Incorporating copying mechanism in sequence-to-sequence learning," *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap.*, vol. 3, pp. 1631–1640, 2016, doi: 10.18653/v1/p16-1154.

[12] A. G. Etemad, A. I. Abidi, and M. Chhabra, "Fine-Tuned T5 for Abstractive Summarization," *Int. J. Performability Eng.*, vol. 17, no. 10, pp. 900–906, 2021, doi: 10.23940/ijpe.21.10.p8.900906.

[13] A. Pal, L. Fan, and V. Igodifo, "Text Summarization using BERT and T5," *Anjali001.Github.Io*, 2022, [Online]. Available: https://anjali001.github.io/Project_Report.pdf

[14] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proc. NAACL-HLT 2016*, San Diego, CA, USA, Jun. 2016, pp. 93–98, doi: 10.18653/v1/N16-1012.

[15] G. Mishra, N. Sethi, A. Loganathan, Y. H. Lin, and Y. C. Hu, "Attention Free BIGBIRD Transformer for Long Document Text Summarization," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 16, no. 2, pp. 210–229, 2024.

[16] A. M. Bakiyeva and T. V. Batura, "Hybrid approach to automatic summarization of scientific and technical texts," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 4, pp. 559–570, 2020.

[17] W. Qiu, Y. Shu, and Y. Xu, "Research on Chinese multi - documents automatic summarizations method based on improved TextRank algorithm and seq2seq," 2021.

[18] S. Bano, S. Khalid, N. M. Tairan, H. Shah, and H. A. Khattak, "Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 9, p. 101739, 2023, doi: 10.1016/j.jksuci.2023.101739.

[19] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[20] V. Gulati, D. Kumar, D. E. Popescu, and J. D. Hemanth, "Extractive article summarization using integrated TextRank and BM25+ algorithm," *Electronics*, vol. 12, no. 2, p. 372, 2023, doi: 10.3390/electronics12020372.

[21] A. F. AlShammari, "Implementation of Keyword Extraction using Term Frequency-Inverse Document Frequency (TF-IDF) in Python," *Int. J. Comput. Appl.*, vol. 185, no. 35, pp. 9–14, 2023, doi: 10.5120/ijca2023923137.

[22] H. Lin, J. Bilmes, and S. Xie, "Graph-based submodular selection for extractive summarization," *Proc. 2009 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2009*, pp. 381–386, 2009, doi: 10.1109/ASRU.2009.5373486.

[23] H. Ahmad, S. A. Rafi, N. Rahman, and K. N. Islam, *Optimizing Abstractive Summarization With Fine-Tuned PEGASUS*, B.Sc. thesis, Dept. of Computer Science and Engineering, Brac University, Bangladesh, Sept. 2023

[24] N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, 2018, doi: 10.1109/ICCUBEA.2018.8697465.

[25] R. Mengi, H. Ghorpade, and A. Kakade, "Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis," *Gt. Lakes Bot.*, pp. 1–7, 2023.

[26] M. Zaheer *et al.*, "Big Bird: Transformers for longer sequences," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 17283–17297, 2020.

[27] N. Ibrahim Altmami and M. El Bachir Menai, "Automatic summarization of scientific articles: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1011–1028, 2022, doi: 10.1016/j.jksuci.2020.04.020.

[28] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," *Nat. Lang. Process. J.*, vol. 7, no. March, p. 100070, 2024, doi: 10.1016/j.nlp.2024.100070.

[29] T. Wang *et al.*, "OPEN A study of extractive summarization of long documents incorporating local topic and hierarchical information," *Sci. Rep.*, pp. 1–13, 2024, doi: 10.1038/s41598-024-60779-z.

[30] S. S. R. Katib and M. H. Abdulameer, "Question Answering System Based on Bideirectional Long-Short-Term Memory (Bilstm)," *Al-Furat J. Innov. Electron. Comput. Eng.*, vol. 3, no. 2, pp. 105–120, 2024, doi: 10.46649/fjiece.v3.2.9a.18.5.2024.

[31] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," *CoNLL 2016 - 20th SIGNLL Conf. Comput. Nat. Lang. Learn. Proc.*, pp. 280–290, 2016, doi: 10.18653/v1/k16-1028.

[32] D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/9365340.

[33] Q. Wang and J. Ren, "Summary-aware attention for social media short text abstractive summarization," *Neurocomputing*, vol. 425, no. xxxx, pp. 290–299, 2021, doi: 10.1016/j.neucom.2020.04.136.

[34] D. Parveen, *A Graph-Based Approach for the Summarization of Scientific Articles*, Ph.D. dissertation, Dept. of Computational Linguistics, Ruprecht-Karls-Universität Heidelberg, 2018.

[35] M. Zhang, G. Zhou, W. Yu, N. Huang, and W. Liu, "A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2022. 2022. doi: 10.1155/2022/7132226.