

Article

Attention Free BIGBIRD Transformer for Long Document Text Summarization

Gitanjali Mishra ¹, Nilambar Sethi ¹, Agilandeewari Loganathan ^{2,*}, Yu-Hsiu Lin ³
and Yu-Chen Hu ⁴

¹ Department of Computer Science and Engineering, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha 765022, India; gitamca3@gmail.com (G.M.); nilambar@giet.edu (N.S.)

² School of Computer Science Engineering and Information Systems (SCORE), Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

³ Graduate Institute of Automation Technology, National Taipei University of Technology, Taipei 1006344, Taiwan; yhlin@ntut.edu.tw

⁴ Department of Computer Science, Tunghai University, Taichung 407224, Taiwan; ychu@thu.edu.tw

* Corresponding author: agila.l@vit.ac.in

Received date: 5 March 2024; Accepted date: 9 May 2024; Published online: 24 May 2024

Abstract: The requirement for creating automatic text summarization systems has dramatically increased as a result of the web's tremendous expansion in textual data and the challenge of finding desired information within this vast volume of data. Transformer-based automatic text summarization using Pre-trained Language models is most attractive in terms of cutting-edge performance and accuracy. Unfortunately, due to their full attention mechanism, one of their key weaknesses is the quadratic memory dependency on the sequence length. Also, the suitability of those transformers for summarizing long documents is another issue. Both these issues can be addressed using a BIGBIRD transformer with a linear computational complexity of $O(n)$. To attain improved accuracy by reducing the redundancy and to improve the similarity among the sentences, a novel attention-free BIGBIRD hierarchical Transformer is introduced in this paper, where the general BIGBIRD involves sparse attention, which is not scalable. Additionally, for handling the long document efficiently, it is crucial to build an effective model using a DistilBART-CNN-12-6 and multi-objective meta-heuristics algorithm that can learn and represent various compositions efficiently by selecting the sentences. These selected sentences from both the meta-heuristics and DistilBART-CNN are given as input for the Attention Free BIGBIRD Transformer to strengthen the summarization ability. Thus, the proposed Attention Free BIGBIRD Transformer using DistilBART-CNN and Meta-Heuristics Algorithm for Long Document Text Summarization (AFBB-LDTS) system achieves a better ROUGE and BLEU score when compared to the related state-of-the-art systems and it is less complex.

Keywords: BIGBIRD; attention-free transformer; multi-objective; meta-heuristics; pre-trained language models; Distil-BART-CNN

1. Introduction

As the internet and big data have expanded, people are becoming increasingly overwhelmed by the volume of data and documents available online. Many academics are motivated by this to create technical methods for automatically summarizing texts. Automatic text synthesis produces summaries that contain key phrases and all pertinent details from the source material [1,2]. As a result, the information is delivered promptly, and the document's original intent is pre-served [3]. The goal of automatic text summarization is to reduce the length of documents' contents into shorter versions of them. Manual text summarization could be time-consuming and expensive [4]. These hitches can be overwhelmed using an Automatic text summarization approach and facilitate the production of the key concepts in a portion of the script contentedly. The current expansion of non-structured textual data in the digital sphere

necessitates the creation of automatic text summary technologies that make it simple for users to conclude them. Implementing summarization can make documents easier to read, save time spent looking up the information again, and allow for the fitting of more information into a given space [5].

According to the International Data Corporation (IDC), the amount of digital data that is transmitted globally each year will increase from 4.4 zettabytes in 2013 to 180 zettabytes in 2025. Because there is so much data floating around in the digital world, algorithms that can automatically condense longer texts and provide precise summaries that effectively convey the intended messages must be developed. Additionally, using text summarization shortens reading sessions, speeds up information research, and expands the amount of information that can fit in a given space. Since the middle of the 20th century, researchers have explored text summarization; Luhn used word frequency diagrams as a statistical tool to explain the topic in public for the first time [6]. There are single-document summaries and multi-document summaries depending on the document count. Meanwhile, the extractive and abstractive outcomes are based on the summary results. A single document generates a summary taken from a single source document, and the content described revolves around the same subject [7]. Multiple documents that address a similar subject were used to create the multi-document summary [8,9]. Extractive summarization refers to creating summaries that are solely made up of content extracted from the source text [10]. Finding the place of the sentence and the frequency of terms in the text were the typical issues that emerged from the extractive summarization research at first [11]. An unsupervised extractive summarization using the rank fusion of multiple features extracted namely topic, semantics, position, and keyword for each sentence. [12]. Recent summarising techniques based on sequence networks fall short of capturing the document's long-range semantics, which are included in its topic vectors. A brand-new technique for extractive document summarization based on topic modelling and word embedding with meta-learning [13]. The goal of DeepSumm is to enhance the quality and accuracy of the summarized text by making use of latent information in the document that has been assessed using topic vectors and sequence networks [14]. A quantum-inspired genetic algorithm-based extractive summarization for multi-documents has also been introduced [15]. The teaching-learning based optimization strategy is utilized to calculate the best weights for the text features, while the final sentence score is determined by a fuzzy inference system that uses a human-generated knowledge base to generate a summary [5,16]. The information extraction (IE) technique was used in the following experiment to address the extraction issue and create a summary with more precise results and greater accuracy. Contrary to extractive summarization, abstractive summarization creates completely new sentences, sometimes known as paraphrases, which provide outlines with words that are never present in the original context. Because they involve considerable natural language processing, abstractive summaries are much more complicated and challenging than extractive summaries [2]. Based on the size of the input document, these abstractive and extractive summarizations are further classified into short document and long documents. The common abstractive summarization algorithms are seq2seq models with RNN [17], attention, copy mechanisms, content selection, pointer-generator methods, and reinforcement learning [18]. By treating the original document's sentences and words as authorities and hubs, the HITS-based attention mechanism fully utilizes the information at the sentence and word levels. It uses Kull-back-Leibler (KL) divergence to refine the attention value which is treated as a novel abstractive summarization method as it produces an enhanced summarization performance [19]. These techniques work well with small documents from high-resource summary datasets like CNN/Daily Mail [20], Gigaword [21], Medical Dataset [22] etc.

However, summarizing lengthy papers with countless tokens is a more pertinent difficulty from a practical standpoint. Existing solutions concentrate on utilizing document structure [23] or mixed model summarization [24], which involves extracting valid sentences without redundant or irrelevant sentences followed by an abstractive summarization. However, a lot of training data is needed for these techniques. When somebody attempts to summarize a lengthy text manually would first comprehend it, then underline the key points, and finally paraphrase it to provide a summary. The pre-trained language models (PTLMs), Reinforcement Learning (RL), Transfer learning (TL), and Deep Learning (DL) are best suitable to handle single documents, multi-documents, and long texts [25]. A hybrid extractive and abstractive text summarization using PTLMs has been proposed recently [26]. This system first creates an extractive summary using several input documents before using it to create an abstractive summary. The first phase deals with redundant information, which is a general issue in multi-document summarization. Redundancy is addressed specifically using the Determinantal Point Process (DPP). The length of the input sequence for the abstractive summarization method is likewise controlled in this step. This action has two results. The first step is to speed up computing. For an abstractive summarizer, the second step is to preserve the key portions of the incoming documents. To assess the quality of the phrases in the extracted summary, a deep submodular network (DSN) is used, and to calculate redundancy,

BERT-based similarities are used. To create two abstractive summaries, the acquired extractive summary is then fed into the pre-trained BART and T5 models. From the two abstractive summaries, one final summary is chosen by looking at the variety of sentences in each summary. These combined Pre-trained language models to overcome the requirement motivate us to develop an efficient summarization model for long documents. Thus the Attention-free BIGBIRD transformer, multi-objective meta-heuristics, and DistilBART-CNN techniques are used in the proposed system to generate a competent summary for the long documents.

The major highlights of the proposed approach are:

- (i) Introduced an efficient Attention-free BIGBIRD (AFBB-LDTS) trans-former for producing long document text summarization in a linear time.
- (ii) The multi-objective meta-heuristic approach used here helps to achieve better sentence scores and reduces redundancy among the sentences.
- (iii) The DistilBART-CNN-12-6 is used to select meaningful sentences.
- (iv) Achieved higher ROUGE and BLEU scores concerning the state-of-the-art (SOTA) systems for a long document.

The organization of the paper is detailed as Section 2 elaborates on the literature survey, Section 3 demonstrates the background information, Section 4 introduces the AFBB-LDTS system, Section 5 illustrates the experimental results, Section 6 presents the comparative analysis and finally, conclusions and future work are drawn in Section 7.

2. Literature Survey

This section elaborates on the existing systems suitable for text summarization. The most attractive solutions for NLP problems in recent days particularly text summarization are handled by deep convolutional neural networks (CNN) which is a form of an encoder and decoder design where the encoder takes the sequences of input text and converts it into a hidden form of text which the decoder can only understand and produces a summary. The first DNN system introduced for English-to-French translation was [27]. The popular deep neural network for text summarization applications is (recurrent neural network (RNN) which works on the sequence input to sequence output. Later, bi-directional RNNs were introduced to keep track of the contexts in both directions [28]. These RNN systems are sequence-to-sequence models that suffer from handling the long-term dependencies of input texts, thus leading to vanishing and exploding gradient problems and producing poor-quality summaries for long documents. Also, parallelism is not possible, and it can be overcome by hierarchical CNNs [29,30], and Optimization Models [31,32]. The basic deep neural network framework described in [33] won't be suitable to handle dependencies for long documents as well as OOV words elongated as Out of Vocabulary. The two-phase ensemble learning approach is introduced for extractive text summarization are also been suitable only for short documents [34].

Next, popular systems for handling long-term dependencies [35] are long short-term memory (LSTM) [36] and gated recurrent unit (GRU) [37]. Comparatively, GRU is simpler than the LSTM systems in terms of complexity, faster training, and execution. These systems can solve vanishing gradient problems but are unable to resolve exploding gradients which can be achieved via gradient clipping [37]. However, these systems handle long-term dependencies and suffer from remembering very long-term documents. A new attempt has been made by [38] called gated recurrent LSTM systems, which can produce extractive and abstractive summaries.

Thus, without using any RNN or CNN nodes, an entirely attention-based architecture was created using the techniques solely for focusing attention by utilizing the Transformer architecture. This transformer architecture provides considerable advancements in the caliber of the outcomes and ROUGE scores have been made more recently [39]. The transformer can be used as an encoder and/or decoder in a variety of models comprising pre-trained language models (PTLMs). Since transformers support treating the input token separately via the self-attention mechanism, it can compare the similarities among those tokens in parallel [40]. As a result, the transformer effectively addressed the issues of RNNs, including their nature of representation in sequences and also the dependencies in long documents [41,42].

Even though transformers outperform the sequence-to-sequence models of RNN and CNN, they suffer from quadratic memory and computational complexities as they require more operations to be performed to handle long text documents. Various attempts have been made to increase the transformer's effectiveness and address its issues with the help of PTLMs such as BERT [43], and Reformer [44], which decreases the complexity in terms of memory and thus the training speed is improved as $O(n \log n)$. As mentioned in [45], the sparse transformer implements sparsity that can support training a deeper network with limited operations and an $O(n \log n)$ complexity preserves $O(n)$ with the help of the self-

attention mechanism. Transformer-XL [46], which uses autoregressive-based architectures, also enables models to comprehend the context and learn dependencies outside a pre-determined duration limit. Recently, a fuzzy-based hybrid model for long documents using extractive and abstractive summarization using Bidirectional GRU has been introduced [47], but it is very complex. To handle the complexity, many metaheuristic techniques have been used in recent years to find the ideal weights for scoring methods or pertinent sentences for summary generation, including Particle Swarm Optimization (PSO) [48], Genetic Algorithm (GA) [49], Harmony Search Algorithm (HSA) [50], Cat Swarm Optimization (CSO) [51], Multicriteria Optimization (MCO) [52], and Jaya [53]. A substantial amount of computational work is needed to tune a large number of regulating factors using these metaheuristic procedures. Furthermore, the algorithm's efficacy is altered by even slight modifications to its settings. As a result, these approaches' performances are wildly inconsistent.

Accordingly, BIGBIRD [4] has been primarily anticipated to handle the longer texts for abstractive automatic text summarization (ATS) by reducing the linear form of dependency, i.e., $O(n)$ times. BIGBIRD is used mainly for an abstractive form of ATS by summarizing long texts and producing state-of-the-art outcomes. This BIGBIRD architecture has two identified issues, namely, (i) scalability and (ii) sparse attention used with a random connection.

From the literature, it is clear that PTLM models namely BERT, Reformer, and Transformer-XL were best suitable for short document text summarization. Moreover, the long document summarization problem can also be handled by the PTLM system but it has a severe complexity issue. Among these, the BIGBIRD system is competent for long documents with the above-mentioned issues of scalability and sparse attention. This motivates us to develop a modified version of BIGBIRD with DistilBART-CNN and optimization for long document summarization. Thus, the modified BIGBIRD is introduced by replacing existing sparse attention with an attention-free transformer, followed by a DistilBART-CNN and meta-heuristic technique that improves the similarity measures and sentence scores by reducing the redundancies among the sentences.

3. Preliminary Concepts

3.1. BIGBIRD Transformer

One of the most effective deep learning models for NLP is transformer-based models, like BERT. Unfortunately, their entire attention mechanism leads to quadratic dependence on sequence length, particularly memory size, which is their major weakness. This quadratic dependence problem is handled by a sparse attention mechanism called BIGBIRD [4] developed by the Google research team which takes only linear time as it dealt only with a sparse attention mechanism instead of full attention in the earlier case. This sparse attention consists of three different connections: sliding, global, and random. These attentions are formed for a given sentence that states "How old are you?"

From Figure 1, it is clear that the number of connections required for full will be more than the sparse level connections. Even though it represents a smaller change in number, it will have a big impact on the computational time.

BIGBIRD uses sparse attention or generalized attention on each layer of the input sequences $X = X_1, X_2, X_3, X_4, \dots, X_n$. For a directed graph D , the generalized attention with vertex set of $V = 1, 2, 3, \dots, n$, Neighbors $N(i)$, then the output vector of the generalized attention is given as,

$$Attn_D(X)_i = x_i + \sum_{n=1}^{H_d} \sigma \left(Q_n(x_i) K_n(X_{N(i)})^T \right) \cdot V_n(X_{N(i)}) \quad (1)$$

where,

Q_n and K_n are the query and key functions, respectively,

V_n is a value function,

σ is a scoring function and

H_d represents the head in numbers.

Also note $X_{N(i)}$ corresponds to the matrix formed by only stacking $\{x_j: j \in N(i)\}$ and not all the inputs.

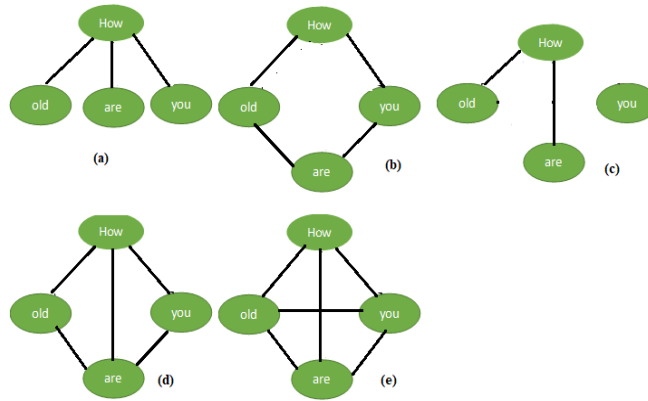


Figure 1. BIGBIRD Sparse Attention vs. Full Attention Mechanism: (a) Global; (b) Sliding; (c) Random; (d) Sparse Connection; (e) Full Connection.

3.2. DistilBART-CNN-12-6

A pre-trained neural language model called “sshleifer/distilbart-cnn-12-6” is based on the DistilBART architecture, which is a distilled form of the BART model. In many natural language processing (NLP) tasks, DistilBART performs similarly to BART while being smaller and faster. To learn language representations, it combines a transformer-based design with a denoising autoencoder. The training configuration of the DistilBART architecture is indicated by the “CNN-12-6” portion of the model name. The model contains 6 encoder-decoder layers and 12 transformer layers. The encoder-decoder layers are utilized for sequence-to-sequence activities like text generation or summarization, while the transformer layers are in charge of encoding the incoming text and creating context-aware representations. The developer who made the model is identified by the suffix “sshleifer” in its name. The “sshleifer/distilbart-cnn-12-6” model is often fine-tuned on a task-specific dataset by modifying the model's parameters and training it on task-specific data to be used for a particular NLP task. The model's pre-trained weights can be utilised to initialize it for the fine-tuning procedure, which can enhance the model's performance on the given task. There are various advantages and drawbacks to using the “sshleifer/distilbart-cnn-12-6” paradigm for text summarization.

3.2.1. Merits

1. Effectiveness: For several natural language processing tasks, including text summarization, the “sshleifer/distilbart-cnn-12-6” model is a state-of-the-art model. It has demonstrated competitive performance on benchmark text summarization datasets, demonstrating its capacity to produce summaries of high quality.
2. Quicker inference: DistilBART models are typically quicker and utilise fewer computational resources than their full-sized equivalents, making them a more effective option for use cases that call for swift inference, such as online summarization applications.
3. Transfer learning: Because the model has already been trained on a huge volume of text data, it can be further optimized for a particular job, like summarization, on a smaller dataset. The use of this transfer learning strategy can produce better results

3.2.2. Demerits

1. Limited training data: The model's ability to generalise to new or unknown material may be constrained by the small quantity of data it was trained on for the summarising task. As a result, it might not function well when applied to datasets that are vastly different from those it was trained on.
2. The model has a maximum output length constraint, which may limit its capacity to provide lengthy summaries. In some instances, this restriction can lead to summaries that are insufficient or incomplete.
3. Possibility of bias: Pre-trained language models may pick up on and reinforce prejudice found in training data. It is essential to be aware of this issue and carefully examine the model's output for any signs of bias or discrimination.

Overall, where speed and efficiency are key considerations, the “sshleifer/distilbart-cnn-12-6” model can be a solid option for text summarising jobs. However, it's critical to assess the model's output and

take into account any flaws or potential biases.

3.3. Attention-Free BIGBIRD Transformer

Even though the BIGBIRD [4] proposed by the Google research team supports long-term dependencies, it suffered from a problem of scalability. This can be overcome by the replacement of the self-attention layer with an Attention-Free Transformer (AFT) proposed by [54], which is a replacement for multi-head attention (MHA) without the need to change other architectural aspects of Transformers. We used AFT-local, where the learned position biases are restricted to a small area while maintaining global connectivity. For any X as an input, AFT first performs a linear transformation to $Q = XW^Q, K = XW^K, V = XW^V$, where $Q, K,$ and V represent Query, Key, and Value respectively, and then performs the following as in [54],

$$Y = f(X)$$

$$Y_i = \sigma_q(Q_i) \odot \frac{\sum_{t'=1}^T \exp(K_{t'} + w_{t,t'}) \odot V_{t'}}{\sum_{t'=1}^T \exp(K_{t'} + w_{t,t'})} \quad (2)$$

where

\odot is the product which is performed element by element

σ_q is the nonlinearity with $\omega \in R^{T \times T}$ as sigmoid.

The above equation is rewritten below to represent the relationship between the MHA and AFT [54]

$$Y_t^i = a_t^i V^i \quad (3)$$

such that
$$a_t^i = \frac{\sigma_q(Q_i) \exp(K^i + w_i)}{\sum_{t'=1}^T \exp(K_{t'}^i + w_{t,t'})}$$

where, $i = 1, 2, 3, 4, 5, \dots, d$ which is the feature dimension of matrix and

$t = 1, 2, 3, 4, 5, \dots, T$ for each position we have an attention vector $a_t^i \in R^T$ for each dimension composed of $Q, K,$ and w .

3.4. Cuckoo Search Optimization

For determining the literal and semantic similarity of the sentences, authors have attempted a metaheuristic method, namely cuckoo search (CS) Optimization, proposed by Yang and Deb (2009) [55]. According to the study, the CS algorithm outperforms other algorithms like GA and PSO in terms of lowering localization error since it has fewer parameters, is simple to implement, is straightforward for novice users to use, and has a far more effective global search capability [56]. However, the CS algorithm has a sluggish rate of convergence, which means more resources are needed to obtain a given level of accuracy. As opposed to other optimization methods, the CS algorithm is dependable and offers a superior solution to the issue [57–60]. The Levy flight concept improves the CS algorithm and is used for determining sentence similarity.

The following analogy has been used by us in our suggested methodology to calculate similarity measures:

1. The new solution is the cuckoo's egg, which illustrates the resemblance in both meaning and literal terms. Additionally, this is utilized to group the sentences.
2. The quality of the eggs for each host nest (sentence subset) is either 1 or 0, indicating whether the sentence subset is chosen for the summary or not.
3. The likelihood that the host bird will find the cuckoo egg it has placed is low (0.30 in our case). It represents eliminating the least important and superfluous sentences (worst sentence subset), which will then be excluded from the further calculation. This likelihood is taken to remain constant.

The cuckoo search optimization is used to identify the semantic similarity between the sentences and to determine the sentence with better sentence scores for further processing. As inspired by [61], to figure out both semantically and literally, we used cosine distance calculation for semantic similarity and Levenshtein distance calculation for similarity in literal.

Thus, the semantic similarity for any two given input sentences ' ISp ' and ' ISq ' is calculated using the cosine similarity as,

$$CS = \text{cos_sim}(ISp, ISq) \quad (4)$$

Similarly, the literal similarity using collinearity of any two given input sentences 'ISp' and 'ISq' is calculated using the Levenshtein distance as,

$$LS = \text{lev_sim}(ISp, ISq) \quad (5)$$

The above equations are considered as two objective functions for the defined cuckoo search optimization as

$$F1 = \text{Max}[\beta (\text{Cos_sim}(ISp, ISq))] \text{ and} \quad (6)$$

$$F2 = \text{Max}[1 - \beta (\text{Lev_sim}(ISp, ISq))] \quad (7)$$

where β is the learning rate of semantic similarity and $1-\beta$ is the learning rate of literal similarity. The overall objective function is represented as,

$$F = \text{Max}(F1, F2) \quad (8)$$

The detailed representation of the Cuckoo search optimization is shown in Figure 2

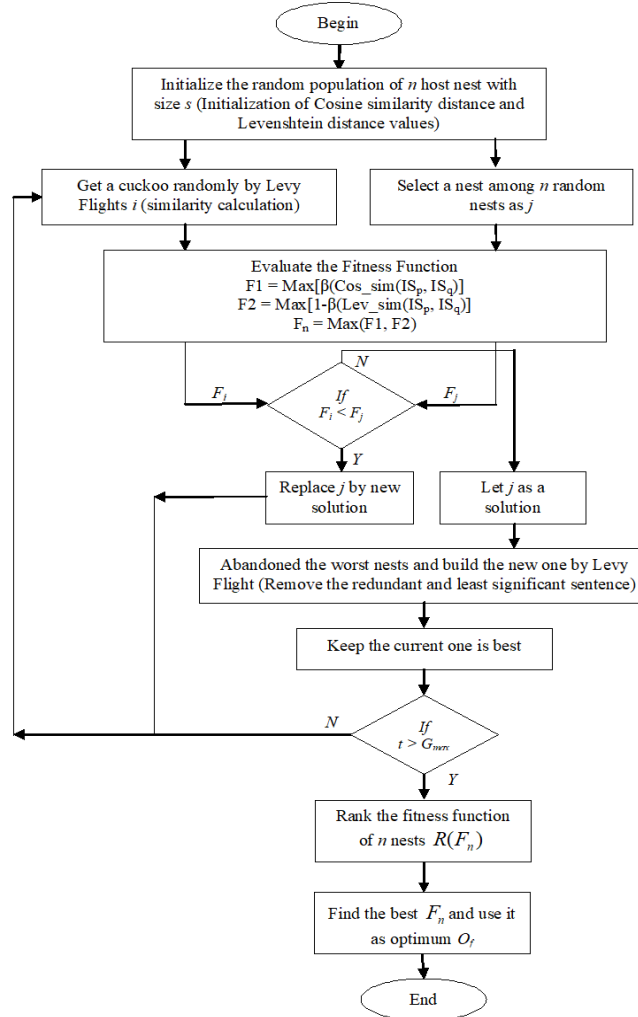


Figure 2. Multi-Objective based Cuckoo Search Optimization for semantic similarity and literal similarity.

4. Proposed AFBB—LDTs Model

This section introduces the enhanced framework for an efficient model of summarization for long as well as short documents. The proposed AFBB-LDTs model involves three phases, namely, (i) Pre-processing, (ii) Extraction of Meaningful sentences, and (iii) Generating the summary. These steps are

illustrated in Figure 3 and also explained in detail as follows:

Step 1: The input documents can be either a short document ‘Ds’ such as CNN/Daily Mail, DUC2002 corpus or a long document ‘Dl’ such as Clinical Texts, arXiv, etc.

Step 2: The input documents are pre-processed using pre-processing techniques such as removal of the stop words, Stemming, Tokenization, and Segmentation.

- Case conversion involves converting the entire text in the input ‘Ds’ or ‘Dl’ document into a lowercase to maintain uniformity throughout the text.
- Removal of stop words concentrates on removing the common words such as ‘a’, ‘an’, and ‘the’ which won’t add any information to the input text.
- Stemming removes the suffix as well as a prefix from the sentences and converts them into basic words.
- Segmentation is used to organize the sentences after it is extracted from the documents.
- Contraction mapping: Contractions are the most common in any online document which deals with contracting a word or groups of words by dipping the letters and supplanting them with an apostrophe. The text summarization gets affected by these contractions because,
 - (i) It won’t be easily understood in their context by conventional systems as it is subjective in use,
 - (ii) It is computationally expensive as it drastically increases the dimensionality of the vectorized text. To resolve this, contraction mapping is applied to map each sentence to its expanded form [60].
- Tokenization is used to extract the words from each sentence, mainly to identify the structure of the character, such as date and time, number, punctuation, etc. In addition, in this approach, we are using a Distilbart tokenizer namely “sshleifer/distilbart-cnn-12-6” to tokenize special tokens needed by the Distilbart models.

$$\begin{aligned} Dstoprem &= \text{Caseconversion}(Ds \text{ or } Dl) \text{ AND } (\text{Stopwords removal}(Dcaseconv)) \text{ AND} \\ Dstem &= \text{Stemming}(Dstoprem) \quad \text{AND} \quad Dseg = \text{Segmentation}(Dstem) \quad \text{AND} \\ Dconmap &= \text{Contraction Mapping} (Dseg) \end{aligned} \quad (9)$$

$$Dpp = Dtoken (\text{AutoTokenizer.from_pretrained}(sshleifer/distilbart-cnn-12-6))$$

Step 3: The pre-processed documents are now passed to two simultaneous operations, namely (i) optimization, and (ii) DistilBART to obtain meaningful sentences or the sentences of the highest scores.

Step 3a: Cuckoo Search Optimization:

Optimized to determine meaningful sentences with the help of semantic similarity and literal similarity using cosine distance and Levenshtein distance, respectively.

$$CS = \text{cos_sim}[Dpp(ISr, ISs)] \quad (10)$$

where *cos sim* represents the cosine similarity distance,

Dpp represents the pre-processed document,

ISr—*r*th sentence of *Dpp*, and

ISs—*s*th sentence of *Dpp*.

Similarly, the literal similarity using collinearity of any two pre-processed input sentences ‘*ISr*’ and ‘*ISs*’ is calculated using the Levenshtein distance as,

$$LS = \text{lev_sim}[Dpp(ISr, ISs)] \quad (11)$$

lev_sim represents the Levenshtein similarity distance,

Dpp represents the pre-processed document,

ISr—*r*th sentence of *Dpp*, and *ISs*—*s*th sentence of *Dpp*

The highest similarity distance is considered the highest score. Thus, the output of an optimization will be the highest sentence score as,

$$\text{Sentence score} = \text{Max}[\text{Max}(\beta(CS), 1 - \beta(LS))] \quad (1)$$

where β represents the learning rate which varies between 0 to 1 (0.25 in our case)

Step 3b: On the other hand, the pre-processed documents namely *Dpp* are fed into the DistilBART model as,

model = AutoModelForSeq2SeqLM.from_pretrained(sshleifer/distilbart-cnn-12-6)

To understand the maximum number of tokens in this model, the following command is used

tokenizer.model_max_length

For calculating the token count limit for longer sentences, the following command is used

$\max([\text{len}(\text{tokenizer.tokenize}(\text{sentence})) \text{ for sentence in sentences}]$

This DistilBART model produces the selected sentences on one hand, on the other hand, the sentences with the highest scores are produced by the Cuckoo Search Optimization technique. These two were the input to the AFT BIGBIRD Transformer.

Step 4: The sentences with the highest sentence score from the Cuckoo Search Optimization and the selected sentences from the DistilBART model are given as input to the AFT-based BIGBIRD, which is suitable for a long document as it supports long-range dependencies. The basic self-attention is described as,

$$\text{Self-Attention}(Q, K, V) = \text{softmax}\left(\frac{Q_i T K_i}{\sqrt{d_K}}\right) V_i \quad (13)$$

where, Q_i = Query, K_i = Key, V_i = Value, d_K = Key dimensions

This modified BIGBIRD replaces the sparse attention by AFT that first performs linear transformation $Q_i = XW_iQ, K_i = XW_iK, V_i = XW_iV$ as in Equations (2) and (3).

The output of this modified BIGBIRD will be a summary generation.

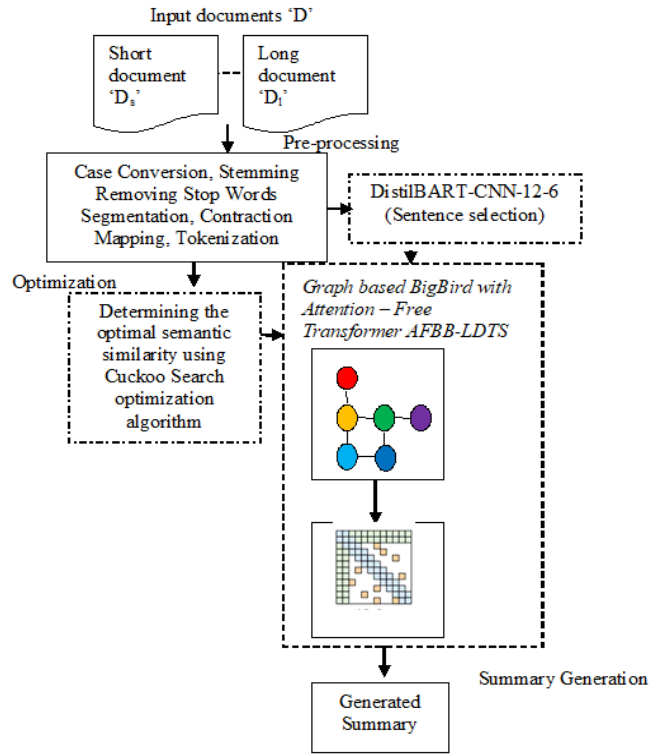


Figure 3. Proposed AFBB-LDTS Model.

5. Experimental Results

To evaluate the performance of the proposed AFBB-LDTS model, it is tested with different short document datasets namely CNN/Daily Mail, BBC News, and DUC, and long document datasets namely arXiv and PubMed. The metrics used for this evaluation are the ROUGE score and BLEU score.

5.1. Dataset Description

The various benchmark datasets used for the evaluation of the proposed AFBB-LDTS model for short documents are CNN / Daily Mail News, BBC News, Document Understanding Conference (DUC), and long documents are arXiv, and PubMed respectively. A detailed description of these datasets is given here.

5.1.1. CNN/Daily Mail

The English-language dataset, particularly the combined form of CNN and Daily Mail is just over 300,000 unique news stories that were authored by journalists for CNN and the Daily Mail. For more instances of data to achieve superior training, the datasets CNN and Daily Mail are united. Although the

initial version was developed for automated reading, comprehension, and abstractive question answering, the current version supports both extractive and abstractive summarization. As per https://huggingface.co/datasets/viewer/?dataset=cnn_dailymail&config=3.0.0 accessed on 10 December 2022, the proposed models also use the average count of the token for the articles and the highlights are 781 and 56, respectively. The 3 major splits of the CNN/Daily Mail dataset are in terms of training, testing, and validation, and the same is projected for Version 3.0.0 in Table 1.

Table 1. CNN / Daily Mail Version 3.0.0 Dataset and its splitting.

Splitting of Dataset	Instances in Each Split
Training	287,113
Testing	11,490
Validation	133,68

5.1.2. BBC News

This dataset was made using a dataset for data categorization from the 2004–2005 work by D. Greene and P. Cunningham [62], which consists of 2,225 documents from the BBC news website relating to the news of five topical divisions (<http://mlg.ucd.ie/datasets/bbc.html>).

5.1.3. DUC

Document Understanding Conference Datasets are generated by the National Institute of Standards and Technology (NIST). The DUC corpus from 2002, 2003, 2004, 2006, and 2007 are used for evaluation and its URL is given as <http://www-nlpir.nist.gov/projects/duc/data> (accessed on 10 December 2022).

5.1.4. PubMed

The PubMed dataset consists of 19,717 scholarly papers on diabetes that have been categorized into one of three categories in the PubMed database. The citation network has 44338 links in it. A word vector from a lexicon with 500 unique terms that is TF/IDF weighted is used to describe each publication in the dataset.

5.1.5. arXiv

There are more than 2 million academic publications in eight subject areas in the arXiv preprint database's complete corpus as of the arXiv annual report 2021. Along with the LaTeX source files, we gather the articles' published PDF versions. For better analysis, the abstract of each paper is collected from the relevant arXiv abstract page. Initial attempts to extract text from the PDFs were made as in [63], however, because PDF parsing is a tough engineering task, it is challenging to construct a clean dataset. Then, as per [64], we decided to parse the approximately 2 million LaTeX source files that were readily available out of the more than 2 million papers. These datasets and their detailed representation in terms of domain, total number of texts, average words, and text size are presented in Table 2.

Table 2. Summary of Benchmark Datasets.

Dataset	Domain	Total Number of Texts	Average Words	Texts Size
CNN dataset	News	90,000	-	Short
Daily Mail	News	197,000	-	Short
BBC	News	2,225	-	Short
DUC 2002	News	567	100 words	Short
DUC 2003	News	350	100 words	Short
DUC 2004	News	500	100 words	Short
DUC 2006	News	1300	250 words	Short
DUC 2007	News	1200	250 words	Short

arXiv	Research	>2 million	-	Long
PubMed	Scientific	19,717	214 words	Long

“-” —Not Available.

5.2. Experimental Setup

The experiments were run on a computer with an NVIDIA Quadro® T550 4GB GDDR6 graphics card and an Intel® Core™ i7-1260P processor with E-cores running at 3.40 GHz and P-cores at 4.70 GHz. The input documents from the dataset are first pre-processed using the above-mentioned pre-processing stages for the short documents news datasets: CNN/Daily Mail, BBC News, DUC, and PubMed. But for a long document such as arXiv, before pre-processing, the collected data is analysed and only the computer science domain articles were collected. The papers belonging to the physics and mathematics domain may contain numerous mathematical expressions in the Introduction section, which might produce noise in language modeling.

Then the Introduction-Abstract pairs are created using an average of 1000 tokens for Introduction and 200 tokens for Abstract. The Introduction-Abstract pair for these articles can be found in the source files, which are used to acquire the Introduction-Abstract pair. Even though it is deterministic, it uses certain effective heuristics, thus, it can typically determine whether single quotes are words or not, whether periods denote sentence borders, etc. Even though the sentence may still contain a few tokens that can come after a sentence-ending character as part of the same phrase, a sentence stops when a sentence-ending character (.,!, or?) is detected that is not grouped with other characters into a token (like for an abbreviation or number). Sentence splitting is a deterministic effect of tokenization (such as quotes and brackets).

5.3. Qualitative Analysis

The qualitative analysis of the proposed summarization model is made by comparing the generated summaries with the outlines written by humans, called human-generated summaries. The results obtained for the various datasets are presented in Figures 4–6. From the figures, it is clear that the proposed summarization model works well for both short and long documents, respectively.

CNN/Daily Mail News Dataset

Actual Text	The government has taken various steps to make India a global drone hub by 2030, Minister of State for Civil Aviation VK Singh told the Rajya Sabha on Monday. The Centre has notified the Liberalised Drone Rules, 2021, and a PLI scheme for drones, published the Drone Airspace Map and UAS traffic management policy framework last year, among other measures taken by the government to boost drone manufacturing and the use of drones in the country, according to the minister.
Ground Truth Summary	Centre Undertakes Series of Reform Measures to Make India Global Drone Hub by 2030, MoS Civil Aviation Says
Generated Summary	Govt plans to brand India as global drone hub.

(a)

Actual Text	The gambling hub of Macao will allow bars, restaurants, salons, gyms and entertainment venues to reopen on Tuesday, August 2, lifting a citywide Covid-19 lockdown that lasted more than a month. Authorities said the decision came after nine consecutive days without reported community infections within the special administration region and more than 14 rounds of mass Covid-19 tests. From Tuesday, people will be required to show proof of a negative Covid-19 test taken within the prior 72 hours to enter most venues.
Ground Truth Summary	China's gambling hub Macao to ease Covid-19 restrictions
Generated Summary	The Macao gambling hub announces reopening with COVID protocols.

(b)

Actual Text	Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Freedman has stabilised Forest since he replaced cult hero Stuart Pearce and the club's owners are pleased with the job he has done at the City Ground. Dougie Freedman is set to sign a new deal at Nottingham Forest. Freedman has impressed at the City Ground since replacing Stuart Pearce in February. They made an audacious attempt on the play-off places when Freedman replaced Pearce but have tailed off in recent weeks. That has not prevented Forest's ownership making moves to secure Freedman on a contract for the next two seasons.
Ground Truth Summary	Nottingham Forest are close to extending Dougie Freedman's contract. The Forest boss took over from former manager Stuart Pearce in February. Freedman has since lead the club to ninth in the Championship.
Generated Summary	Freedman has stabilised Forest since he replaced cult hero Stuart Pearce and the club's owners are pleased with the job he has done at the City Ground.

(c)

Actual Text	Liverpool target Neto is also wanted by PSG and clubs in Spain as Brendan Rodgers faces stiff competition to land the Fiorentina goalkeeper, according to the Brazilian's agent Stefano Castagna. The Reds were linked with a move for the 25-year-old, whose contract expires in June, earlier in the season when Simon Mignolet was dropped from the side. A January move for Neto never materialised but the former Atletico Paranaense keeper looks certain to leave the Florence-based club in the summer. Neto rushes from his goal as Juan Iturbe bears down on him during Fiorentina's clash with Roma in March. Neto is wanted by a number of top European clubs including Liverpool and PSG, according to his agent. It had been reported that Neto had a verbal agreement to join Serie A champions Juventus at the end of the season but his agent has revealed no decision about his future has been made yet. And Castagna claims Neto will have his pick of top European clubs when the transfer window re-opens in the summer, including Brendan Rodgers' side. 'There are many European clubs interested in Neto, such as for example Liverpool and Paris Saint-Germain,' Stefano Castagna is quoted as saying by Gazzetta TV. Fiorentina goalkeeper Neto saves at the feet of Tottenham midfielder Nacer Chadli in the Europa League. 'In Spain too there are clubs at the very top level who are tracking him. Real Madrid? We'll see. 'We have not made a definitive decision, but in any case he will not accept another loan move elsewhere.' Neto, who represented Brazil at the London 2012 Olympics but has not featured for the senior side, was warned against joining a club as a No 2 by national coach Dunga. Neto joined Fiorentina from Atletico Paranaense in 2011 and established himself as No1 in the last two seasons. British Airways has a seat pitch of 31 inches, while easyJet has 29 inches, Thomson's short haul seat pitch is 28 inches, and Virgin Atlantic's is 30-31.
Ground Truth Summary	Fiorentina goalkeeper Neto has been linked with Liverpool and Arsenal. Neto joined Fiorentina from Brazilian outfit Atletico Paranaense in 2011. He is also wanted by PSG and Spanish clubs, according to his agent.
Generated Summary	Liverpool target Neto is also wanted by PSG and clubs in Spain as Brendan Rodgers faces stiff competition to land the Fiorentina goalkeeper, according to the Brazilian's agent Stefano Castagna. Neto is wanted by a number of top European clubs including Liverpool and PSG, according to his agent.

(d)

Figure 4. Generated Summary of our proposed AFBB-LDTS for CNN News Dataset (a)–(d) Sample News.

Russia's war in Ukraine

Over 6.6 million people have been internally displaced in Ukraine because of the hostilities, the International Organization for Migration said Wednesday.

A survey conducted by the IOM last month showed that 15% of Ukraine's population has been displaced.

At the same time, 5.5 million people who were previously displaced have returned home, most to Kyiv city and region, as well as Kharkiv, Odessa and Chernihiv regions.

The organization said many of the people who have been displaced are facing economic hardship. It said that 60% of those who were employed before displacement have lost their jobs and as many as 9% have had no income since the outbreak of the full-scale war in late February.

With the approaching colder months, many are worried about their living conditions, the IOM said. As many as 44% said they needed help with repairs and more than one fourth feared needing to leave their current accommodation due to insufficient heating ahead of winter.

Generated Summary:
Over 6.6 million people have been internally displaced in Ukraine because of the hostilities, the IOM says. At the same time, 5.5 million people who were previously displaced have returned home. Many of the people who have been displaced are facing economic hardship.

(a)

	TEST DATASET	Introduction
1	NaN	Acnesol Gel is an antibiotic that fights bacte...
2	NaN	Ambrodil Syrup is used for treating various re...
3	NaN	Augmentin 625 Duo Tablet is a penicillin-type ...
4	NaN	Azithral 500 Tablet is an antibiotic used to t...
5	NaN	Alkasol Oral Solution is a medicine used in th...
...
996	NaN	Azapure Tablet belongs to a group of medicines...
997	NaN	Arimidex 1mg Tablet is used alone or with oth...
998	NaN	Arpimune ME 100mg Capsule is used to prevent y...
999	NaN	Amlodac CH Tablet is a combination medicine us...
1000	NaN	Angizem CD 120 Capsule ER is used to treat ang...

Generated Summary 1: Acnesol Gel is used to treat acne, which appears as spots or pimples on your face, chest or back.

Generated Summary 2: Ambrodil Syrup is used for treating various respiratory tract disorders associated with excessive mucus.

Generated Summary 16: Ambrodil-S Syrup is used to treat coughs and colds.

Generated Summary 24: Allegra 180mg Tablet belongs to a group of medicines called antihistamines.

(b)

Figure 5. Generated Summary of our proposed AFBB-LDTS for (a) General News, and (b) Medicine Dataset.

Introduction – Abstract Pair https://arxiv.org/abs/1602.06023	In this work, we model abstractive text summarization using Attentional Encoder Decoder Recurrent Neural Networks, and show that they achieve state-of-the-art performance on two different corpora. We propose several novel models that address critical problems in summarization that are not adequately modeled by the basic architecture, such as modeling key-words, capturing the hierarchy of sentence-to-word structure, and emitting words that are rare or unseen at training time. Our work shows that many of our proposed models contribute to further improvement in performance. We also propose a new dataset consisting of multi-sentence summaries, and establish performance benchmarks for further research. Abstractive text summarization is the task of generating a headline or a short summary consisting of a few sentences that captures the salient ideas of an article or a passage. We use the adjective 'abstractive' to denote a summary that is not a mere selection of a few existing passages or sentences extracted from the source, but a compressed paraphrasing of the main contents of the document potentially using vocabulary unseen in the source document. This task can also be naturally cast as mapping an input sequence of words in a source document to a target sequence of words called summary. In the recent past, deep-learning based models that map an input sequence into another output sequence, called sequence-to-sequence models, have been successful in many problems such as machine translation (Bahdanau et al., 2014), speech recognition (Bahdanau et al., 2015) and video captioning (Venugopalan et al., 2015). In the framework of sequence-to-sequence models, a very relevant model to our task is the attentional Recurrent Neural Network (RNN) encoder-decoder model proposed in Bahdanau et al. (2014), which has produced state-of-the-art performance in machine translation (MT), which is also a natural language task. Despite the similarities, abstractive summarization is a very different problem from MT. Unlike in MT, the target (summary) is typically very short and does not depend very much on the length of the source (document) in summarization. Additionally, a key challenge in summarization is to optimally compress the original document in a lossy manner such that the key concepts in the original document are preserved, whereas in MT, the translation is expected to be loss-less. In translation, there is a strong notion of almost one-to-one word level alignment between source and target, but in summarization, it is less obvious. We make the following main contributions in this work: (i) We apply the off-the-shelf attentional encoder-decoder RNN that was originally developed for machine translation to summarization, and show that it already outperforms state-of-the-art systems on two different English corpora. (ii) Motivated by concrete problems in summarization that are not sufficiently addressed by the machine translation based model, we propose novel models and show that they provide additional improvement in performance. (iii) We propose a new dataset for the task of abstractive summarization of a document into multiple sentences and establish benchmarks.
Title	Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond
Generated Summary	This systems proposes novel models that can address the serious problems in summarization that cannot be tolerated by the basic architecture, namely keywords modelling, seizing the sentence-to-word structure hierarchy and rare or unseen words need to be omitted during the time of training. This system also develop performance goals for future study and suggest a new dataset made up of multi-sentence summaries.

Figure 6. Generated Summary of our proposed AFBB-LDTS for a Long Document.

From the qualitative results shown in Figures 4–6, we infer that the proposed AFBB-LDTS model produces meaningful summarization results for various benchmark datasets, namely CNN/DailyMail, BBC, General News, PubMed Dataset, and Long documents (arXiv) when compared to its ground truth summary.

5.4. Quantitative Analysis

5.3.1. ROUGE

The common quantitative metric for measuring the summarization model is the Recall Oriented Understudy for Gisting Evaluation (ROUGE) [65]. It is an evaluation tool that is co-selection based, which involves counting overlapping units between the candidate summary and other human-generated summaries, such as the n-gram (ROUGE – N), Longest Common Subsequence (ROUGE – L), and Weighted Longest Common Subsequence (ROUGE – W), the quality of the created summary is assessed using this metric. ROUGE – N is a formal n-gram recall measure between the generated summary and the ground truth summary (N = 2 in our studies). If N = 1, then ROUGE – 1 is obtained, called Unigrams and Bigrams can be obtained with N = 2. The proposed AFBB-LDTS ROUGE score is depicted in Table 3.

$$ROUGE - N = \frac{\sum_{Se \{Ground Truth Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{Se \{Ground Truth Summaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (14)$$

where,

n represents the n-gram’s length

$gram_n$ is the maximum co-occurring n-grams in the generated summary.

$Count_{match}(gram_n)$ is the maximum number of co-occurring n-grams in a set of reference summaries,

The count is the total n-grams in the reference summaries.

5.3.2. BLEU

The measure BLEU (BiLingual Evaluation Understudy) is used to evaluate machine-translated text automatically. The resemblance of the machine-translated text to a collection of excellent reference translations is gauged by the BLEU score, which ranges from zero to one. The proposed AFBB-LDTS BLEU score is given in Table 4.

Table 3. ROUGE Score of Proposed Summarization model on the benchmark dataset.

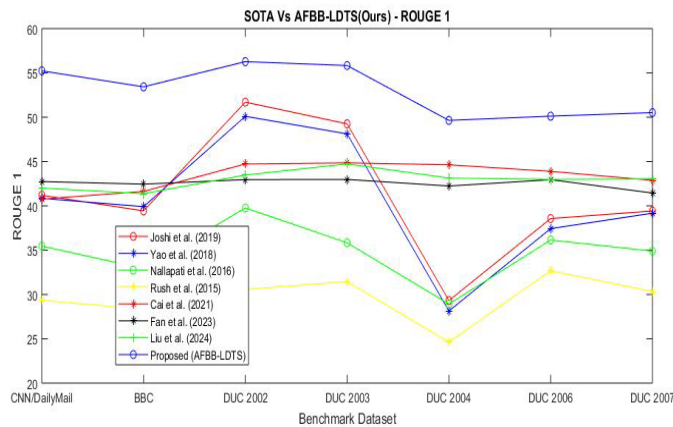
Dataset	ROUGE 1	ROUGE 2	ROUGE L
CNN/Daily Mail	55.23	29.56	52.45
BBC	53.43	27.74	50.25
DUC 2002	56.27	29.99	53.93
DUC 2003	55.83	28.65	52.76
DUC 2004	49.64	23.92	47.14
DUC 2006	50.12	24.86	47.95
DUC 2007	50.52	25.94	48.64
arXiv	48.93	23.24	46.88
PubMed	49.86	24.78	48.92

Table 4. BLEU Score of Proposed Summarization model on the benchmark dataset.

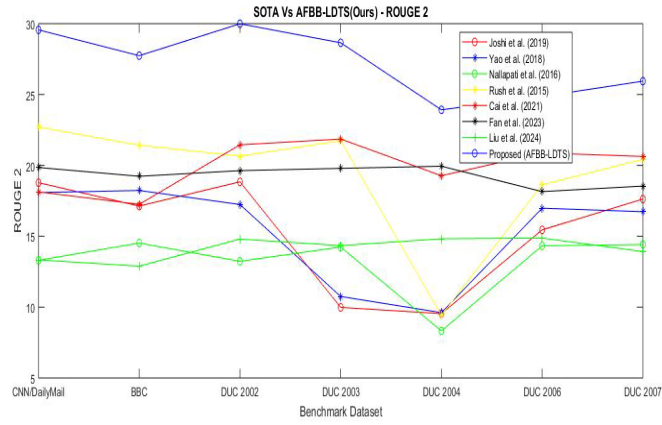
Dataset	BLEU
CNN/Daily Mail	55.67
BBC	54.82
DUC 2002	61.12
DUC 2003	59.78
DUC 2004	51.36
DUC 2006	53.27
DUC 2007	53.75
arXiv	50.59
PubMed	51.68

6. Comparative Analysis

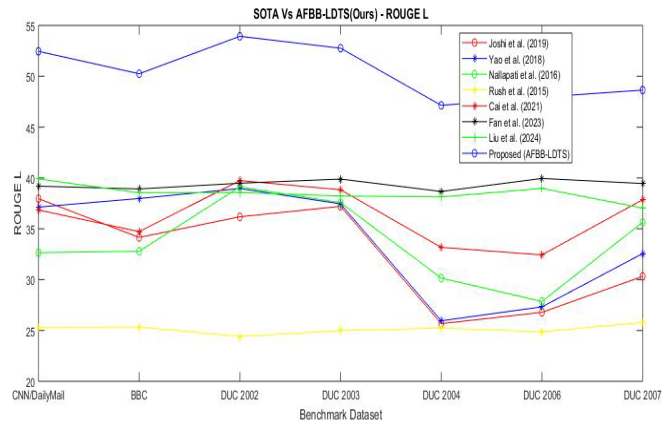
For a fair comparison, we have tested our model with the state-of-the-art (SOTA) systems using the benchmark datasets for both short documents, namely CNN / Daily Mail, BBC news, and DUC, are (i) abstractive text summarization using sequence to sequence RNNs by Nallapati et al. [20], (ii) abstractive text summarization using neural attention by Rush et al. [21], (iii) automatic extractive text summarization using autoencoders by Joshi et al. [66], (iv) abstractive text summarization using dual encoding by Yao et al. [67], (v) abstractive text summarization using attentional neural model [19], (vi) MFMMR-BertSum by Fan et al. [68], and (vii) Improved TextRank Algorithm and K-Means Clustering by Liu et al. [69]. For long documents, the benchmark datasets namely arXiv and PubMed are compared with the SOTA systems, (i) Pre-trained BigBird-Pegasus by Manzil Zaheer et al. [4], (ii) Discourse-Aware based pointer generator model by Cohan et al. [23], and BERT and BiGRU deep extractive approach by Bano et al. [70]. The comparative analysis of ROUGE—1, ROUGE—2, ROUGE—L, and BLEU scores of state-of-the-art systems (SOTA) with the proposed AFBB-LDTS are presented in Figure 7. From Figure 7, it is inferred that the Proposed AFBB-LDTS system outperforms the SOTA systems because of its attention-free model.



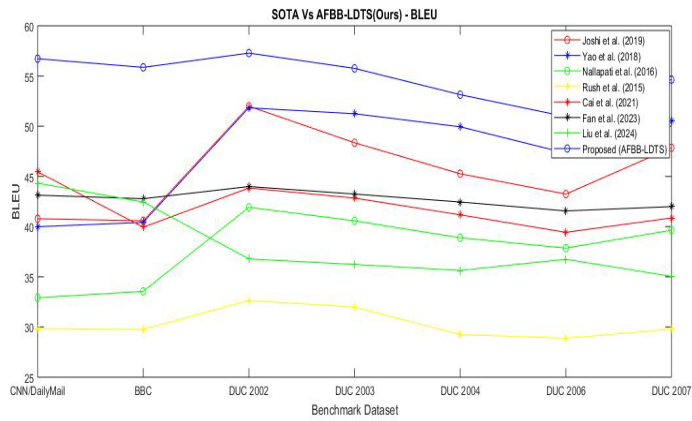
(a)



(b)



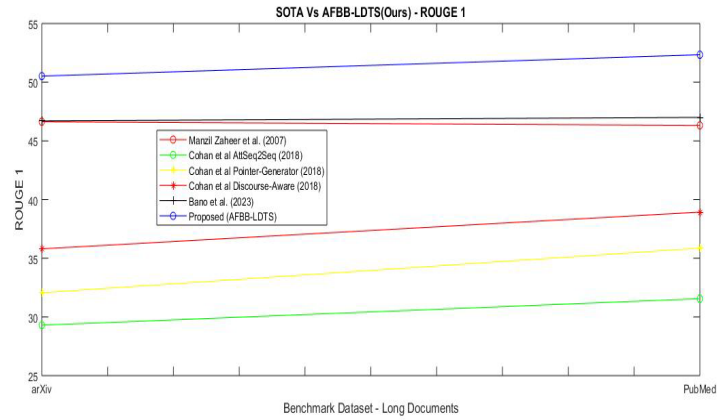
(c)



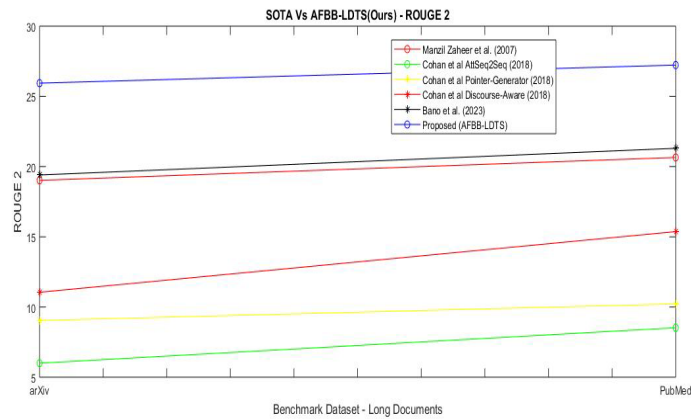
(d)

Figure 7. Proposed AFBB-LDTS—Short documents; (a) ROUGE—1, (b) ROUGE—2, (c) ROUGE—L, and (d) BLEU score. Source: [19–21, 66–69].

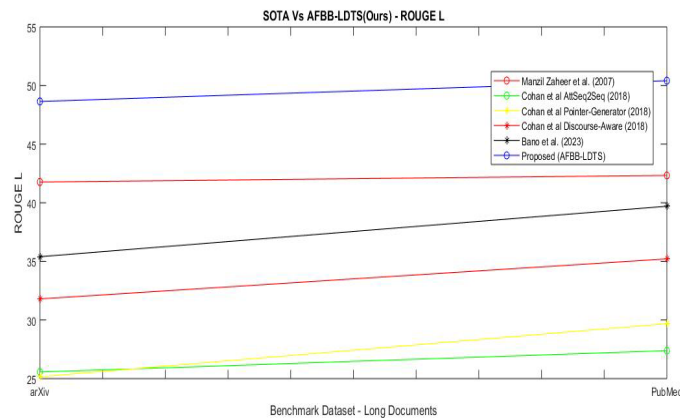
From Figure 7, it is clear that the proposed model's ROUGE score is higher when compared to the state-of-the-art (SOTA) systems [19–21], and [66–69] in terms of ROUGE–1, ROUGE–2, ROUGE–L, and BLEU score this is because of the Distil-BART CNN and Attention Free BigBird Transformer. From Figure 8, it is evident that the proposed AFBB-LDTS model outperforms the other SOTA systems [4,23,70] for long documents because the proposed attention-free model achieves both time efficiency and long-term dependency and thus ensures long document summarization with better ROUGE and BLEU scores.



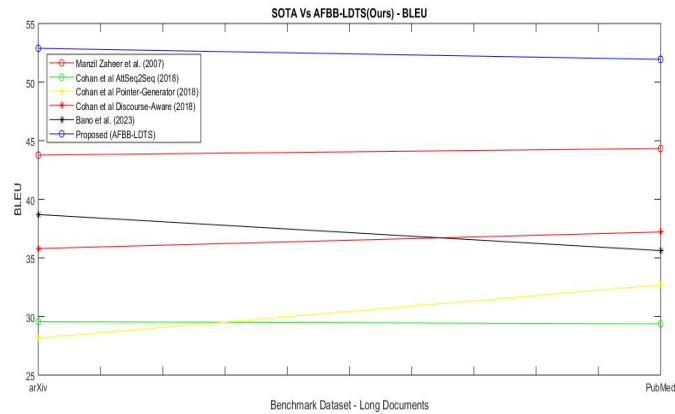
(a)



(b)



(c)



(d)

Figure 8. Proposed AFBB-LDTS—Long documents; (a) ROUGE—1, (b) ROUGE—2, (c) ROUGE—L, and (d) BLEU score ([4,23,70]).

7. Conclusions and Future Work

In this research, we offer a unique attention-free BIGBIRD hierarchical Trans-former to achieve increased accuracy by minimizing the redundancy and improving the similarity among the phrases, while the general BIGBIRD involves sparse attention, which is not scalable. Furthermore, to effectively handle the lengthy document, a multi-objective meta-heuristics method and DistilBART-CNN-12-6 must be used to create an efficient model that can learn and represent different compositions by choosing sentences. To improve the summarization capabilities, the Attention Free BIGBIRD Transformer receives these chosen sentences from both the DistilBART-CNN and the meta-heuristics. Thus, the proposed AFBB-LDTS outstrips the other transformer models in terms of computational time and handles the long document efficiently. The proposed AFBB-LDTS system achieves a better ROUGE—1, ROUGE—2, ROUGE—L, and BLEU score as 55.23, 29.26, 52.45, and 56.72 respectively when compared to the state-of-the-art systems and is also less complex. AFBB-LDTS won't be suitable for long multi-documents. The future work can be in handling long multi-documents, especially for the biomedical domain with the help of lightweight and efficient shifted hierarchical transformers.

Author Contributions

The following statements should be used “Conceptualization, Gitanjali Mishra and Nilambar Sethi; methodology, Gitanjali Mishra; software, Gitanjali Mishra; validation, Gitanjali Mishra, Nilambar Sethi, Agilandeewari L, Yu-Hsiu Lin, and Yu-Chen Hu.; writing—original draft preparation, Gitanjali Mishra; writing—review and editing, Nilambar Sethi, Agilandeewari L, Yu-Hsiu Lin, and Yu-Chen Hu.; All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare there is no conflict of interest.

Data Availability Statement

The links for the publicly available datasets are given as: **CNN/DailyMail** https://huggingface.co/datasets/viewer/?dataset=cnn_dailymail&config=3.0.0, (accessed on 10 Dec 2022). **BBC News** (<http://mlg.ucd.ie/datasets/bbc.html>). (accessed on 10 Dec 2022)). **Document Understanding Conference (DUC) Datasets** <http://www-nlpir.nist.gov/projects/duc/data>. (accessed on 10 Dec 2022)).

References

1. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Ko-chut, K. “Text summarization techniques: A brief survey”. arXiv arXiv:1707.02268. 2017.
2. Gambhir, M., & Gupta, V. “Recent automatic text summarization techniques: A survey”. Artificial Intelligence Review, 47(1), pp. 1-66, 2017.
3. Murad, M. A. A., & Martin, T. “Similarity-based estimation for document summarization using Fuzzy sets”. International Journal of Computer Science and Security, 1(4), pp. 1-12. 2007.

4. Zaheer, M.; Guruganesh, G.; Dubey, A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. "Big Bird: Transformers for Longer Sequences". arXiv arXiv2007.140622020.
5. Appalla, P., Kuthadi, V. M., & Marwala, T. "An efficient educational data mining approach to support e-learning". *Wireless Networks*, 23, 1011-1024. 2017.
6. Luhn, H. P. "The automatic creation of literature abstracts". *IBM Journal of Research and Development*, 2(2), pp. 159-165. 1958.
7. Radev, D. R., Blair-Goldensohn, S., & Zhang, Z. (2001, September). Experiments in single and multidocument summarization using MEAD. In *First Document Understanding Conference* (p. 1A8). 2001.
8. Qiang, J. P., Chen, P., Ding, W., Xie, F., & Wu, X. (2016). Multi-document summarization using closed patterns. *Knowledge-Based Systems*, 99, 28-38. 2016
9. John, A., Premjith, P. S., & Wilscey, M. (2017). Extractive multi-document summarization using population-based multicriteria optimization. *Expert Systems with Applications*, 86, 385-397. 2017.
10. Khan, A. Salim, N., 2014. A review on abstractive summarization. *Journal of Theoretical Applied Inf. Technology*, 59, 64 – 72. 2014
11. Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development*, 2(4), 354-361. 1958
12. Joshi, A., Fidalgo, E., Alegre, E., & Alaiz-Rodriguez, R. "RankSum—An unsupervised extractive text summarization based on rank fusion". *Expert Systems with Applications*, 200, pp. 116846. 2022.
13. Vo, S. N., Vo, T. T., & Le, B. "Interpretable extractive text summarization with meta-learning and Bi-LSTM: A study of meta learning and explainability techniques". *Expert Systems with Applications*, 245, 123045. 2024.
14. Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. "DeepSumm: Exploit-ing topic models and sequence to sequence networks for extractive text summarization". *Expert Systems with Applications*, 211, pp.118442. 2023.
15. Mojriani, M., & Mirroshandel, S. A. "A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA". *Expert systems with applications*, 171, pp. 114555.51, 2021.
16. Verma, P., Verma, A., & Pal, S. "An approach for extractive text summarization using fuzzy evolutionary and clustering algorithms". *Applied Soft Computing*, 120, pp. 108670. (2022).
17. Zhang, Y.; Li, D.; Wang, Y.; Fang, Y.; Xiao, W. Abstract Text Summarization with a Convolutional Seq2seq Model. *Applied. Science*, 9, 1665. 2009. <https://doi.org/10.3390/app9081665>.
18. Wu, Y., & Hu, B. (2018, April). Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1). 2018.
19. Cai, X, Shi, K, Y. Jiang, L. Yang, S. Liu, "HITS-based attentional neural model for abstractive summarization", *Knowl.Based Syst.* 222 (2021) 106996, <http://dx.doi.org/10.1016/J.KNOSYS.2021.106996>.
20. Nallapati, R., Zhou, B., Gulcehre, C., & Xiang, B. "Abstractive text summarization using sequence-to-sequence RNNs and beyond". arXiv preprint arXiv:1602.06023. 2016.
21. Rush, A. M., Chopra, S., & Weston, J. "A neural attention model for abstractive sentence summarization". arXiv preprint arXiv:1509.00685. 2015.
22. Agilandeewari L, Akash Dagar, Deepthi A, Arangasakthivel R, Automatic Text Summarization for Medical Dataset - An Analysis, ISDA 2023.
23. Cohan, A., Deroncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. "A discourse-aware attention model for abstractive summarization of long documents". arXiv preprint arXiv:1804.05685. 2018.
24. Liu, L., Lu, Y., Yang, M., Qu, Q., Zhu, J., & Li, H. (2018, April). Generative adversarial network for abstractive text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1). 2018
25. Alomari, A., Idris, N., Sabri, A. Q. M., & Alsmadi, I. "Deep reinforcement and transfer learning for abstractive text summarization: A review". *Computer Speech & Language*, 71, 101276. 2022.
26. Ghadimi, A., & Beigy, H. "Hybrid multi-document summarization using pre-trained language models". *Expert Systems with Applications*, 192, 116292. (2022).
27. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1724–1734. <https://doi.org/10.3115/v1/d14-1179>.
28. Al-Sabahi, K.; Zuping, Z.; Kang, Y. Bidirectional attentional encoder-decoder model and bidirectional beam search for abstractive summarization. arXiv Prepr. arXiv1809.066622018.
29. Shi, T., Keneshloo, Y., Ramakrishnan, N., Reddy, C.K., 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM Trans. Data Sci.* 2, 1–37.
30. Kim, S.-E.; Kaibalina, N.; Park, S.-B. A Topical Category-Aware Neural Text Summarizer. *Appl. Sci.* 2020, 10, 5422. <https://doi.org/10.3390/app10165422>
31. Jain, A.; Arora, A.; Morato, J.; Yadav, D.; Kumar, K.V. Automatic Text Summarization for Hindi Using Real Coded Genetic Algorithm. *Appl. Sci.* 2022, 12, 6584. <https://doi.org/10.3390/app12136584>
32. Mohsin, M.; Latif, S.; Haneef, M.; Tariq, U.; Khan, M.A.; Kadry, S.; Yong, H.-S.; Choi, J.-I. Improved Text Summarization of News Articles Using GA-HC and PSO-HC. *Appl. Sci.* 2021, 11, 10511. <https://doi.org/10.3390/app112210511>
33. Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2, pp. 3104–3112.
34. Mishra, G., Sethi, N., & Agilandeewari, L. (2023, March). "Two Phase Ensemble Learning Based Extractive Summarization for Short Documents". In *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)* (pp. 129-142). Cham: Springer Nature Switzerland.

35. Chandar, S., Sankar, C., Vorontsov, E., Kahou, S.E., Bengio, Y., 2019. Towards non-saturating recurrent units for modelling long-term dependencies. In: Proc. AAAI Conf. Artif. Intell., 33, pp. 3280–3287. <https://doi.org/10.1609/aaai.v33i01.33013280>.
36. Hochreiter, S., 1997. LSTM can solve hard long time lag problems. *Adv. Neural Inf. Process. Syst.* 473–479.
37. Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In: *Proceedings of the International Conference on Machine Learning*. PMLR, pp. 1310–1318.
38. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. Language modeling with gated convolutional networks. In: *34th Int. Conf. Mach. Learn. ICML 2017*, 2, pp. 1551–1559.
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., 2017. Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 5999–6009. 2017-Decem
40. Cha, Y., & Lee, Y. “Advanced sentence-embedding method considering token importance based on explainable artificial intelligence and text summarization model”. *Neuro-computing*, 564, 126987. 2024.
41. Choukhi, H.; Alsuhaibani, M. Deep Transformer Language Models for Arabic Text Summarization: A Comparison Study. *Appl. Sci.* 2022, 12, 11944. <https://doi.org/10.3390/app122311944>
42. Siragusa, G.; Robaldo, L. Sentence Graph Attention for Content-Aware Summarization. *Appl. Sci.* 2022, 12, 10382. <https://doi.org/10.3390/app122010382>
43. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. B. (2019). Pre-training of deep bi-directional transformers for language understanding In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics, 4171–86.
44. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The Efficient Transformer. *arXiv Prepr. arXiv2001.044512020*.
45. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating Long Sequences with Sparse Transformers. *arXiv Prepr. arXiv1904.105092019*.
46. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R., 2020. Trans-former-XL: attentive language models beyond a fixed-length context. In: *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 2978–2988. <https://doi.org/10.18653/v1/p19-1285>.
47. Mishra, G., Sethi, N., & Agilandeswari, L. (2023, March). “Fuzzy Bi-GRU Based Hybrid Extractive and Abstractive Text Summarization for Long Multi-documents”. In *Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022)* (pp. 153-166). Cham: Springer Nature Switzerland. 2022.
48. Verma P and Om H. “A variable dimension optimization approach for text summarization”. In: *Proceedings of the Conference on Harmony Search and Nature Inspired Optimization Algorithms*. Springer, pp. 687–696, 2019.
49. Meena Y K and Gopalani D, “Evolutionary algorithms for extractive automatic text summarization”. *Procedia Computer Science* 48: 244–249, 2015.
50. Shareghi E and Hassanabadi L S, “Text summarization with harmony search algo-rithm-based sentence extraction”. In: *Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, ACM, pp. 226–231, 2008.
51. Rautray R and Balabantaray R C. “Cat swarm optimization based evolutionary framework for multi document summarization”. *Physica A: Statistical Mechanics and its Applications* 477: 174–186, 2017.
52. Ansamma J, Premjith P S, and Wilsy M. “Extractive multi-document summarization using population-based multicriteria optimization”. *Expert Systems with Applications* 86: 385–397. 2017.
53. Verma P and Om H. “Collaborative ranking-based text summarization using a metaheuristic approach”. In *Proceedings of the Conference on Emerging Technologies in Data Mining and Information Security*. Springer, pp. 417–426, 2019.
54. Zhai, S., Talbott, W., Srivastava, N., Huang, C., Goh, H., Zhang, R., and Susskind, J. “An Attention Free Transformer.” *arXiv*, Sep. 2021. <https://arxiv.org/pdf/2105.14103.pdf>
55. Yang XS, Deb S (2009) Cuckoo search via Levy flights. In: *Proc. of World Congress on Nature and Biologically Inspired Computing (NaBIC 2009)*, pp. 210–214
56. Shehab, M., Khader, A. T., & Al-Betar, M. A. (2017). A survey on applications and variants of the cuckoo search algorithm. *Applied Soft Computing*, 61, 1041-1059.
57. Xin-She Y (2014) *Cuckoo search and firefly algorithm theory and applications*. Springer
58. Civicioglu P, Besdok E (2013) A conceptual comparison of the cuckoo search, particle swarm optimization, differential evolution, and artificial bee colony algorithms. *Artif Intell Rev* 39(4):315–346
59. Prabukumar, M., Agilandeswari, L., & Ganesan, K. (2019). An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier. *Journal of ambient intelligence and humanized computing*, 10(1), 267-293.
60. Agilandeswari, L., & Ganesan, K. (2018). RST invariant robust video watermarking algorithm using quaternion curvelet transform. *Multimedia Tools and Applications*, 77(19), 25431-25474.
61. Wang, S., Zhao, X., Li, B., Ge, B., & Tang, D. (2017, June). Integrating extractive and abstractive models for long text summarization. In *2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 305-312). IEEE.
62. Greene, D., & Cunningham, P. (2006, June). Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 377-384).
63. Dai, A. M., Olah, C., & Le, Q. V. “Document embedding with paragraph vectors”. *arXiv preprint arXiv:1507.07998*. 2015.
64. Moirangthem, D. S., & Lee, M. “Abstractive summarization of long texts by rep-resenting multiple compositionality with temporal hierarchical pointer generator network”. *Neural Networks*, 124, 1-11. 2020.
65. Lin, C. -Y., & Hovy, E. “Automatic evaluation of summaries using n-gram co-occurrence statistics”. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational*

Linguistics on Human Language Technology, Volume 1 (pp. 71–78). Association for Computational Linguistics. 2003.

66. Joshi A, E. Fidalgo, E. Alegre, L. Fernández-Robles, “SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders”, *Expert Systems with Applications* 129, pp. 200–215. 2019.
67. Yao K, L. Zhang, D. Du, T. Luo, L. Tao, Y. Wu, “Dual encoding for abstractive text summarization”, *IEEE Trans. Cybern.* 50 (3), pp.985–996. (2018).
68. Fan, J., Tian, X., Lv, C., Zhang, S., Wang, Y., & Zhang, J. “Extractive social media text summarization based on MFMMR-BertSum”. *Array*, 20, 100322. 2023.
69. Liu, W., Sun, Y., Yu, B., Wang, H., Peng, Q., Hou, M., ... & Liu, C. “Automatic Text Summarization Method Based on Improved TextRank Algorithm and K-Means Clustering”. *Knowledge-Based Systems*, 111447. 2024.
70. Bano, S., Khalid, S., Tairan, N. M., Shah, H., & Khattak, H. A. “Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach”. *Journal of King Saud University-Computer and Information Sciences*, 35(9), 101739. 2023.




Author Biographies



Gitanjali Mishra is a research scholar at GIET University, Gunupur, Odisha. She completed her M.Tech from Berhampur University, Odisha in 2011. Her main areas of research interest are NLP and Deep Learning. She has around fifteen years of teaching experience.

Nilambar Sethi received his Ph.D. in Computer Science from Berhampur University, Odisha, in 2013 and his Magister degree from Utkal University, Odisha, in 2004. He is currently working as an Associate Professor at the Computer Sciences Department of GIET University, Odisha. His research interests include Machine learning, Data Science, and Natural language processing. In addition, he has served as a Technical Program Committee member of some international conferences and workshops. Dr. Sethi is also an active reviewer of some renowned international journals.



Agilandeewari Loganathan    completed her Ph.D. and working as a Professor in the Department of Software Systems and Engineering, School of Computer Science Engineering and Information Systems (SCORE), VIT Vellore. She received her Bachelor’s degree in Information Technology and Master’s in Computer Science and Engineering from Anna University in 2005 and 2009 respectively. She has around 20 years of teaching experience and published 70+ papers in peer-reviewed reputed journals. Her reputed publications include research articles in peer-reviewed journals namely *Expert Systems with Applications*, *IEEE Access*, *Journal of Ambient Intelligence and Humanized Computing*, *Multimedia Tools and Applications*, and *Journal of Applied Remote Sensing* indexing at Thomson Reuters with an average impact factor of 5. She is a peer reviewer in journals including *IEEE Access*, *Pattern Recognition*, *International Journal of Remote Sensing*, *Array*, *Artificial Intelligence Review*, *Informatics in Medicine Unlocked*, *Neurocomputing*, *Computers*, and *Electrical Engineering*, *Journal of King Saud University– Computer and Information Sciences*, *IET Review*, and *Journal of Engineering Science and Technology (JESTEC)*. She also published about 13 engineering books as per Anna University Syllabus. She is a lifetime member of the Computer Society of India. Her areas of interest include image and video watermarking, image processing, neural networks, cryptography fuzzy logic, machine learning, IoT, information-centric networks, and remote sensing. She can be contacted by email: agila.l@vit.ac.in.



Yu-Hsiu Lin working in the Graduate Institute of Automation Technology, National Taipei University of Technology, Taiwan. He has 1477 citations overall. He published 40+ papers in peer-reviewed reputed journals. His areas of interest include AI, IoT, Fog/Edge Computing, and Smart Grid.



Yu-Chen Hu working as a Professor in the Department of Computer Science at Tunghai University, Taiwan. He has 4904 citations overall. He published 100+ papers in peer-reviewed reputed journals. He is an Editor-in-Chief and peer reviewer for peer-reviewed journals. His areas of interest include Data Compression, Image Processing, Information Hiding, Information Security, and Deep Learning.