Exploring Abstractive Text Summarization: Methods, Dataset, Evaluation, and Emerging Challenges

Yusuf Sunusi, Nazlia Omar, Lailatul Qadri Zakaria
Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia 43600

***Abstract*—The latest advanced models for abstractive summarization, which utilize encoder-decoder frameworks, produce exactly one summary for each source text. This systematic literature review (SLR) comprehensively examines the recent advancements in abstractive text summarization (ATS), a pivotal area in natural language processing (NLP) that aims to generate concise and coherent summaries from extensive text sources. We delve into the evolution of ATS, focusing on key aspects such as encoder-decoder architectures, innovative mechanisms like attention and pointer-generator models, training and optimization methods, datasets, and evaluation metrics. Our review analyzes a wide range of studies, highlighting the transition from traditional sequence-to-sequence models to more advanced approaches like Transformer-based architectures. We explore the integration of mechanisms such as attention, which enhances model interpretability and effectiveness, and pointer-generator networks, which adeptly balance between copying and generating text. The review also addresses the challenges in training these models, including issues related to dataset quality and diversity, particularly in low-resource languages. A critical analysis of evaluation metrics reveals a heavy reliance on ROUGE scores, prompting a discussion on the need for more nuanced evaluation methods that align closely with human judgment. Additionally, we identify and discuss emerging research gaps, such as the need for effective summary length control and the handling of model hallucination, which are crucial for the practical application of ATS. This SLR not only synthesizes current research trends and methodologies in ATS, but also provides insights into future directions, underscoring the importance of continuous innovation in model development, dataset enhancement, and evaluation strategies. Our findings aim to guide researchers and practitioners in navigating the evolving landscape of abstractive text summarization and in identifying areas ripe for future exploration and development.**

Keywords*—*Abstractive text summarization; systematic literature review; natural language processing; evaluation metrics; dataset; computation linguistics

I. Introduction

In recent times, there has been a significant increase in demand for the utilization of data from diverse sources such as scientific articles, medical records, and social media platforms. The essence of text summarization lies in condensing a lengthy document into a concise, coherent summary. Automatic text summarization techniques fall into two main categories: extractive and abstractive. Extractive summarization methods pull specific words and phrases directly from the original text [1], while abstractive summarization creates new words and phrases that may not appear in the source document, mimicking the way humans summarize [2, 3]. The goal of abstractive summarization is to craft a condensed version of the original text without losing its original meaning [4]. This process involves generating summaries through a process akin to human thought, demanding high capabilities in characterizing, understanding, and producing text from models.

The Sequence-to-Sequence (Seq2Seq) model, which utilizes Recurrent Neural Networks (RNN) including variants like simple RNN, LSTM, and GRU, is a popular choice for abstractive summarization. These models employ an encoder-decoder framework [5-7] where the encoder converts the input text into a context vector, and the decoder then uses this vector to create an abstractive summary. However, Seq2Seq models, especially those based on RNNs, tend to miss crucial information from the original, lengthy texts and may generate redundant content, particularly in tasks involving long documents or sequences [5]. The addition of an attention mechanism to the encoder-decoder structure has shown success in abstractive text summarization [8]. Furthermore, Blekanov et al. [9] explored transformer-based models like LongFormer and T5, comparing them with BART in experiments with real-world Reddit data. Incorporating synthetic data alongside real data, a method often used in machine translation for resource-scarce situations to enhance translation quality, has been effective. Specifically, employing an iterative back-translation strategy, where back-translation systems are trained repeatedly, has shown promising results in improving machine translation. Furthermore, researchers have explored variations of this technique, including multi-round iterative back-translation and adversarial back-translation to further enhance translation quality.

A thorough review of the literature is essential for the progress of research in text summarization. Syed et al. [10] provides an in-depth analysis of key aspects of abstractive summarization, covering trends, general methodologies, tools, and evaluation techniques in this area. Additionally, a survey by Nazari and Mahdavi [11] delves into various approaches and methods utilized in text summarization, categorizing them into statistical, machine learning, semantic-based, and swarm intelligence approaches. Other scholarly articles focus on narrower topics such as specific summarization techniques [12, 84], methodologies employed [13], and evaluation strategies [14].

Unlike other review studies, this review paper offers an up-to-date extensive explanation of all the ins and outs of abstractive text summarization and highlights current key gaps and challenges in the domain which will help other researchers in identifying and addressing them. The process of conducting a research on abstractive text summarization poses significant challenges, particularly for researchers who are newly acquainted with the domain. This complexity arises from several critical factors that demand rigorous scholarly effort and methodological precision. Firstly, the interdisciplinary nature

of abstractive text summarization, which intersects with fields such as natural language processing, artificial intelligence, and computational linguistics, necessitates a comprehensive understanding of diverse theoretical frameworks and technological advancements. Secondly, the rapid pace at which new research and innovations are introduced in this area means that scholars must continuously update their knowledge base, making it difficult to establish a definitive set of studies for review. Thirdly, the task of abstractive summarization itself, which involves generating new text that captures the essence of source documents, presents unique challenges in evaluating the quality and relevance of research findings.

These factors contribute to the intricacy of reviewing literature in this field, requiring dedicated effort to navigate the vast and evolving body of knowledge. Thus, the aims of this study are outlined as follows: a) To pinpoint, examine, and categorize the research topics and trends within abstractive summarization, b) To offer a comprehensive review of the different methods used in abstractive summarization, including the strengths and weaknesses of the prevalent techniques, c) To provide a succinct description of the commonly employed and newest methodologies in this domain, d) To detail the necessary pre-processing steps and the features utilized, e) To explore the challenges encountered in abstractive summarization, addressing both the resolved and outstanding issues, f) To analyze the evaluation strategies and datasets that have been applied, and g) To suggest directions for future research in text summarization. In pursuit of broadening research prospects in this area, this study employs a Systematic Literature Review (SLR) to achieve a more structured, quantifiable exploration with a wider and more varied range of topics, as highlighted by Shaffril et al. [15]. SLR boasts advantages over traditional review methods through its scientific approach and systematic execution, aiming to reduce bias and ensure transparent, verifiable outcomes. The conducted activities include detailing the review process in Section 2, explaining the derived results in Section 3, discussing the responses to research questions raised in Section 2 in Section 4, and concluding the study in Section 5.

II. Methods

A comprehensive search strategy was implemented to identify and collect as many relevant studies as possible on the subject. The search methodology was developed based on the following research questions:

1) What are the current trends in abstractive text summarization?
2) What datasets are used for developing abstractive text summarization models?
3) What are the evaluation metrics used to measure the quality of abstractive summaries?
4) What are the current challenges emerging in the domain?

In this stage, the research questions were deconstructed into distinct concepts to generate search terms and databases and to identify additional sources for exploration. Consequently, search terms were derived from the research questions:

1) Abstractive Text Summarization techniques/algorithms/models.
2) Abstractive summarization datasets.
3) Evaluation metrics.

The initial search string was created using these search terms and then refined by incorporating alternative terms, including synonyms and variant spellings. For this Systematic Literature Review (SLR), the following libraries were utilized to select pertinent literature that addresses the research questions:

1) Web of Science
2) Semantic Scholar
3) Springer
4) ACM
5) Elsevier
6) IEEE

The identified search terms were employed to locate conference and journal articles within six electronic databases, as shown in Fig. 1. Searches were confined to titles, abstracts, and keywords, with the exception of Google Scholar, where only titles were searched. Additionally, the reference sections of pertinent studies were reviewed for cross-citations. Secondary studies, including existing literature surveys, were also acquired.

Numerous searches were performed; however, the search criteria that yielded the most relevant results were:

1) ”Text Summarization” AND ”Abstractive”
2) ”Text Summarization” AND ”Abstractive” AND (“Techniques” OR “Methods”)
3) ”Text Summarization” AND ”Abstractive” AND (”Evaluation Metrics” OR ”Metrics”)

Further filtration of the search results was done by:

1) Removing duplicate documents
2) Devising inclusion and exclusion criteria to identify related papers and discard those that are irrelevant.
3) Performing quality assessment to ensure that papers with high quality were included.

A. Inclusion Criteria

The following are the criteria for paper inclusion:

1) Papers utilizing abstractive technique.
2) Papers based on abstractive summarization that include evaluation metrics.
3) The most recent version/edition of the paper.
4) Papers published between 2018 and 2023.

B. Exclusion Criteria

The following are the criteria for paper inclusion:

1) Papers utilizing abstractive technique.
2) Papers based on abstractive summarization that include evaluation metrics.
3) The most recent version/edition of the paper.
4) Papers published between 2018 and 2023.

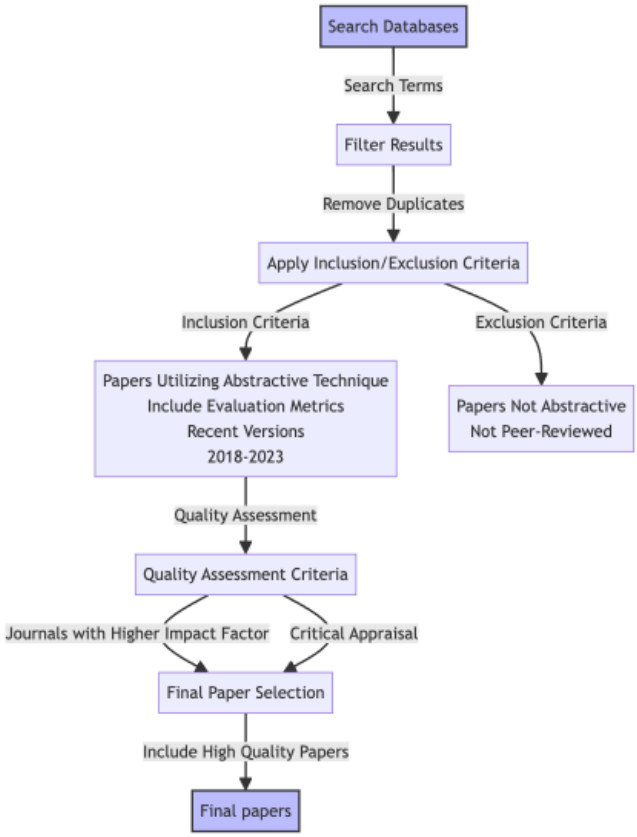

Fig. 1. Search procedure.

C. Quality Assessment

The quality assessment criteria employed in this review were meticulously developed based on the comprehensive framework proposed by Smith et al. [16]. The framework involves a systematic and rigorous set of criteria aimed at ensuring the validity, reliability, and applicability of the research findings. The following are the criteria for assessing the quality of selected journals:

- Journals with higher impact factors.
- Each study will be assessed based on the following critical appraisal criteria:
  - What type of research question is being asked?
  - Was the study design appropriate for the research question?
  - Did the study methods address the most important potential sources of bias?
  - Was the study performed according to the original protocol?
  - Is the study question relevant?
  - Does the study add anything new?
  - Do the data justify the conclusions?
  - Are there any conflicts of interest?
  - Does the study test a stated hypothesis?
  - Were the statistical analyses performed correctly?

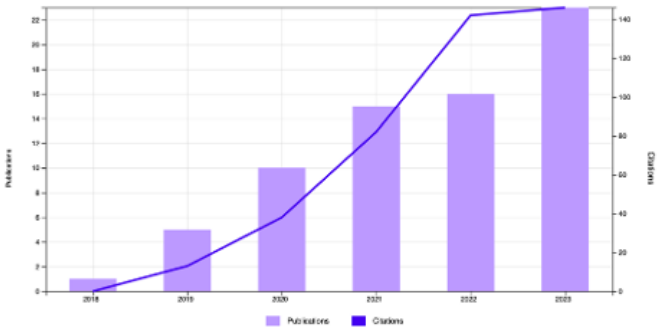

Fig. 2. Publications and citations from results.

D. Data Extraction

Upon searching using keywords "abstractive text summarization (All Fields) and natural language processing (Author Keywords) or Text Summarization (Author Keywords) and abstractive (All Fields) and 2023 or 2022 or 2021 or 2020 or 2019 or 2018 (Publication Years) and Article (Document Types)", a total of 72 journals and publications were ultimately derived. As seen in Fig. 2, research in the field of abstractive summarization is steadily growing each year with 2023 having the highest publications in recent years.

III. Abstractive Text Summarization

Abstractive text summarization is a process where a new, condensed version of a text is generated, capturing the essential messages and meaning of the original content. Unlike extractive summarization, which selects and rearranges specific sentences or phrases from the source text, abstractive summarization involves understanding and interpreting the text to produce summaries that are not necessarily found verbatim in the source. Studies have explored various methodologies, including the use of pre-trained models for better context understanding and the application of novel training paradigms to improve the summarization of specific languages or domains [17, 18]. Table I provides a summary of the categories and terms to be discussed in this review.

A. Encoder Decoder Architecture

Choosing an encoder-decoder framework offers various design options for our encoder and decoder, including traditional RNN/LSTM/GRU, bidirectional RNN/LSTM/GRU, Transformer, BERT/GPT-2 models, or the more recent BART architecture. In the model described by Fan et al. [2], both the encoder and decoder are built as deep convolutional networks. They begin with a layer for word embedding, followed by a series of convolutions alternating with Gated Linear Units (GLU). The decoder is linked to the encoder via attention mechanisms that compute a weighted average of the encoder's outputs. These weights are determined based on the current state of the decoder, enabling it to focus on the most pertinent sections of the input document for generating the subsequent token. Zhang et al. [19] introduced a novel generative model leveraging a convolutional seq2seq framework, which includes a copying mechanism to address rare or unseen words. Furthermore, this model integrates a hierarchical attention mechanism to simultaneously consider both key words and key sentences.

TABLE I. CATEGORIES COVERED IN THIS STUDY

Category	Terms
Encoder-Decoder Architecture	• RNN/LSTM/GRU • Bi-RNN/ Bi-LSTM/ Bi-GRU • Transformer • BERT/GPT2 • BART
Mechanisms	• Attention • Copying • Coverage • Pointer generator
Training and Optimization	• Word Level Training • Sequence Level Training • Document-Level Training • Sentence-Level Training • Transfer Learning • Reinforcement Learning
Dataset	• CNN/DailyMail • New York Times • Gigaword • DUC 2004
Evaluation	• ROUGE • BLEU • METEOR

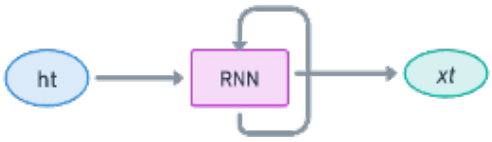

Fig. 3. RNN architecture.

1) Recurrent Neural Network (RNN): Recurrent Neural Networks (RNNs) are a class of neural networks that excel in processing sequential data. Unlike traditional feedforward neural networks, RNNs have loops in their architecture, allowing information to persist [20]. This unique feature makes RNNs particularly suitable for tasks where context and sequence matter, such as language modeling and behavior analysis in social media [21]. The core functionality of an RNN can be captured by the following formula:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \quad (1)$$

Here, h_t is the current hidden state, W_{hh} is the weight associated with the previous hidden state h_{t-1}, x_t is the current input, and W_x is the weight associated with the current input. The $\tanh$ function is an activation function that helps to normalize the output. This formula represents how RNNs process information over time. The current hidden state h_t is a function of the previous hidden state and the current input, allowing the network to maintain a form of 'memory' of past inputs. This is crucial for tasks that require understanding the sequence of data, such as text processing.

Fig. 3 illustrates the looping structure of RNNs, where the output of a layer is fed back into the same layer as input. This loop enables the network to pass information across time steps, effectively remembering previous inputs and using this memory to influence the output.

RNNs have been pivotal in advancing abstractive text summarization. Their ability to handle sequential data makes them ideal for this task, where understanding the context and flow

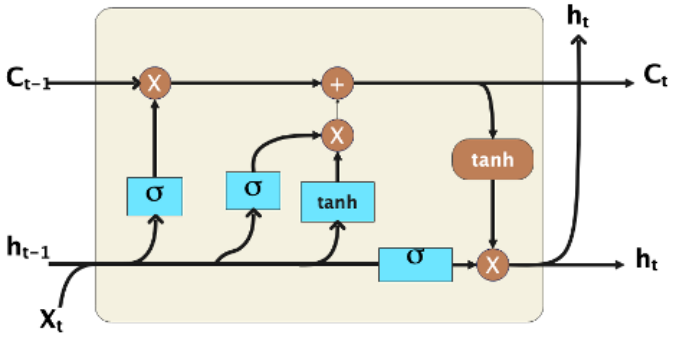

Fig. 4. LSTM architecture.

of a text is crucial for generating coherent summaries. RNNs, especially when combined with techniques like Long Short-Term Memory (LSTM) networks, have significantly improved the performance of abstractive summarization systems [22, 23]. These networks can capture long-range dependencies in text, allowing for more accurate and contextually relevant summaries.

2) Long-Short Term Memory (LSTM): Long Short-Term Memory (LSTM) networks are an advanced type of Recurrent Neural Networks (RNNs), designed to address the challenge of learning long-term dependencies. LSTMs are particularly known for their ability to overcome the vanishing gradient problem commonly encountered in traditional RNNs [20]. LSTM networks introduce a more complex computational unit called a cell, which includes mechanisms known as gates [87]. These gates control the flow of information, allowing the network to retain or forget information selectively. The core operations within an LSTM cell can be summarized with the following formulas:

- Forget Gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- Input Gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- Cell State Update: $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- Final Cell State: $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
- Output Gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
- Hidden State: $h_t = o_t * \tanh(C_t)$

Here, σ represents the sigmoid function, $\tanh$ is the hyperbolic tangent function, W and b are weights and biases, and $*$ denotes element-wise multiplication. These equations represent the LSTM's ability to regulate the flow of information. The forget gate decides what information to discard from the cell state, while the input gate updates the cell state with new information. An illustration of this process is presented in Fig. 4. The output gate then determines what part of the cell state should be outputted to the next layer or used as the hidden state for the next time step.

Fig. 4 illustrates the internal structure of an LSTM cell, showing the input x_t, the previous hidden state h_{t-1}, the cell state C_t, and the gates that regulate the flow of information within the cell. See et al. [24] introduced a sequence-to-sequence LSTM model with attention mechanisms to improve the quality of abstractive summaries, demonstrating significant

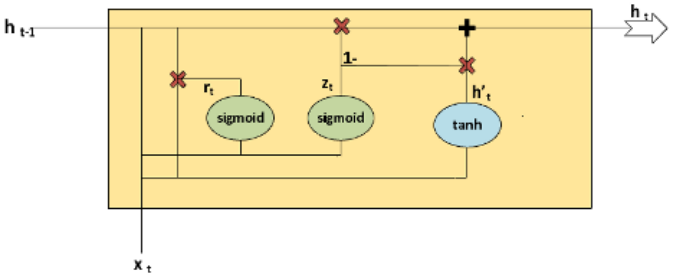

Fig. 5. GRU architecture.

advancements over previous techniques. The model's capacity to deal with long texts and its flexibility in generating novel textual content have made LSTMs a cornerstone in the development of abstractive summarization models, propelling the field towards more human-like summarization capabilities.

3) Gated Recurrent Unit (GRU): GRUs, introduced by Cho et al. [25], are designed to adaptively capture dependencies of different time scales in a sequence. Gated Recurrent Units (GRUs) are a type of recurrent neural network (RNN) architecture that has gained popularity due to their efficiency and effectiveness, especially in sequence modeling tasks. GRUs simplify the LSTM architecture and often provide comparable performance. They consist of two gates: the update gate and the reset gate. These gates help the model decide how much of the past information needs to be passed along to the future. The operations within a GRU can be summarized with the following formulas:

- Update Gate: $z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$
- Reset Gate: $r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$
- Candidate Hidden State: $\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t] + b)$
- Final Hidden State: $h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$

In this context, σ represents the sigmoid function, $\tanh$ is the hyperbolic tangent function, W and b are weights and biases, and $*$ denotes element-wise multiplication. These equations represent how GRUs can selectively update their hidden state. The update gate z_t decides how much of the past information needs to be passed to the future, while the reset gate rt determines how much of the past information to forget. The candidate hidden state h t is a combination of the current input and the past hidden state, modulated by the reset gate. The final hidden state ht is then a blend of the old state and the new candidate state, as governed by the update gate.

Fig. 5 illustrates the internal structure of a GRU cell, showing the input x_t, the previous hidden state h_{t-1}, the update and reset gates, and the final hidden state h_t. GRUs, with their architecture designed to mitigate the vanishing gradient problem common in traditional RNNs, allow for better retention of information over longer sequences which is crucial for summarization tasks. Recent research has further explored the integration of GRUs into sophisticated neural network models to enhance abstractive summarization. For instance, Rehman et al. [26] developed an attentive GRU-based encoder-decoder model, demonstrating the efficacy of GRUs

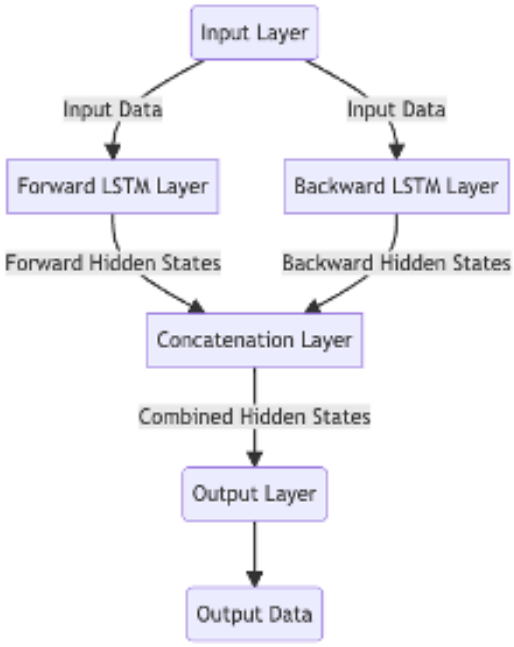

Fig. 6. Bi-LSTM structure.

in producing summaries that are not only concise but also capture the essence of the original text with high fidelity. This research underscores the versatility of GRUs in dealing with diverse linguistic structures and their capacity to improve the summarization process, making them an indispensable tool in the field of natural language processing.

4) Bi-Directional RNN/LSTM/GRU: Bi-directional Recurrent Neural Networks (Bi-RNNs), including their variants like Bi-LSTM (Bi-directional Long Short-Term Memory) and Bi-GRU (Bi-directional Gated Recurrent Unit), are advanced neural network architectures that process data in both forward and backward directions. This bidirectional approach allows the networks to have both backward and forward information about the sequence at every time step [27].

Fig. 6 illustrates a bi-directional architecture where each time step receives inputs from two sides: one from the beginning of the sequence to the current time step and the other from the end of the sequence to the current time step. This dual input mechanism allows the network to preserve information from both past and future states, enhancing its predictive accuracy.

In abstractive text summarization, the context of the entire text is crucial for generating coherent and relevant summaries. Bi-directional models, by considering both preceding and following contexts, can capture the nuances of language more effectively. This results in summaries that are not only concise, but also maintain the essence and flow of the original text. Preethi et al. [27] developed an abstractive summarizer using Bi-LSTM, demonstrating its capability to produce precise and coherent summaries without the redundancy issues often encountered in simpler models.

5) Transformer: The Transformer is a type of neural network architecture that has garnered significant attention, particularly in the domain of natural language processing

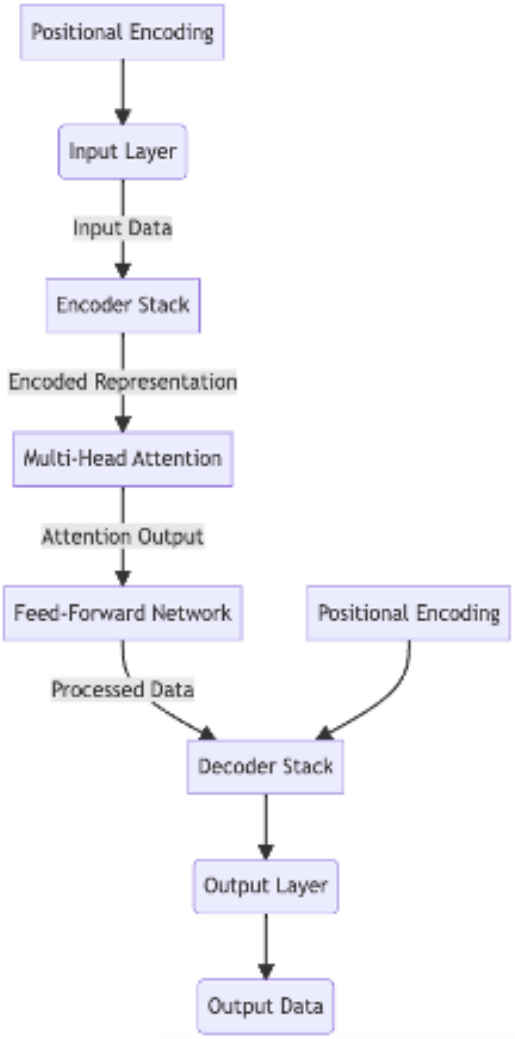

Fig. 7. Transformer architecture.

(NLP). Introduced by Vaswani et al. [28], the Transformer model is known for its reliance on self-attention mechanisms, which allow it to process input data in parallel and capture complex dependencies in sequences. Transformers have been applied in various domains, demonstrating their versatility and effectiveness. Fig. 7 illustrates the core components of the Transformer architecture, including the encoder and decoder stacks, each comprising multiple layers of self-attention and feed-forward neural networks. The self-attention mechanism allows the model to weigh the influence of different parts of the input data, enabling it to capture long-range dependencies.

a) Text-to-Text Transfer Transformer (T5): The Text-to-Text Transfer Transformer (T5) model, an encoder-decoder Transformer implementation, has been pivotal in the advancement of abstractive text summarization. This model's ability to convert all NLP problems into a unified text-to-text format, as outlined by Raffel et al. [29], allows for seamless application across a wide range of text summarization tasks. T5's architecture, incorporating self-attention mechanisms, layer normalization, and a dense layer with a softmax output, facilitates the generation of coherent and contextually relevant summaries from extensive text inputs. In the realm of summarization, T5 has been utilized to push the boundaries of abstractive text summarization, offering significant improvements over traditional models. For instance, Itsnaini et al. [30] leveraged T5 in the context of the Indonesian language, showcasing its effectiveness through high evaluation scores despite challenges in achieving optimal abstraction. Further research by Lubis et al. [18] introduced an approach by combining T5 with Bayesian optimization to enhance text summarization. Additionally, the study on Arabic news summarization by Ismail et al. [17] utilized a T5-based approach, achieving state-of-the-art performance and illustrating T5's capacity for language-specific applications. These examples highlight the versatility and efficiency of the T5 model in handling diverse and complex summarization tasks. By leveraging pre-trained models like T5, researchers can address the inherent challenges of abstractive summarization, such as maintaining factual accuracy and coherence, across various languages and domains.

6) Bidirectional Encoder Representations from Transformer-Generative Pre-Trained Transformer (BERT-GPT): Bidirectional Encoder Representations from Transformer (BERT), when combined with Generative Pre-Trained Transformer (GPT) which excels in generating coherent and contextually relevant text, becomes particularly effective for abstractive text summarization. For instance, a study by Darapaneni et al. [31] explored the use of BERT and GPT-2 models for summarizing research articles related to COVID-19. They found that while BERT models performed well for extractive summarization, there was room for improvement in abstractive summarization, which was addressed by using GPT-2 models. The combination of these models helped in creating more accurate and comprehensive summaries. In another study, Baykara and Güngör [32] utilized pre-trained sequence-to-sequence models, including BERT and GPT, for Turkish abstractive text summarization. They demonstrated that these models could generate high-quality summaries by addressing challenges such as saliency, fluency, and semantics. Kieuvongngam et al. [33] also leveraged BERT and GPT-2 for summarizing COVID-19 medical research articles. Their approach provided abstractive and comprehensive summaries based on keywords extracted from the original articles, showcasing the effectiveness of these models in processing complex medical texts.

The integration of Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) models has shown promising results in the field of abstractive text summarization. These models leverage the strengths of both BERT's deep bidirectional understanding and GPT's powerful generative capabilities. The combination of BERT and GPT models brings together the best of both worlds – deep contextual understanding and advanced text generation capabilities. This synergy is particularly beneficial for abstractive text summarization, where the goal is to generate summaries that are not only concise but also retain the essence and context of the original text. By leveraging BERT's ability to understand nuanced context and GPT's proficiency in generating coherent text, these models can create summaries that are both informative and readable, making them highly suitable for summarizing complex and lengthy documents.

7) Bidirectional and Autoregressive Transformers (BART): Bidirectional and Auto-Regressive Transformers (BART)'s architecture, which includes a bidirectional encoder (like BERT) and an autoregressive decoder (like GPT), enables it to understand the context of a text deeply and generate summaries that are both fluent and informative [34]. BART has emerged as a powerful tool in the field of abstractive text summarization. It combines the benefits of both autoencoding and autoregressive approaches, making it particularly effective for generating coherent and contextually accurate summaries. Baykara and Güngör [32] utilized BART for Turkish abstractive text summarization where they showed that BART, along with other pre-trained sequence-to-sequence models, could generate high-quality summaries by addressing challenges such as saliency, fluency, and semantics. BART's effectiveness in abstractive text summarization stems from its ability to understand and reconstruct the input text accurately. By pre-training on a large corpus of text with various noising and denoising tasks, BART learns to correct errors, fill in missing information, and rephrase sentences, which are essential skills for summarization. When fine-tuned on summarization tasks, BART can generate summaries that not only capture the essential points of the original text, but also maintain a natural and coherent narrative flow. This makes BART an ideal choice for summarizing complex texts across various domains including news articles, scientific papers, and legal documents.

B. Mechanisms

Mechanisms serve as additional features integrated into the fundamental neural encoder-decoder structure, aimed at resolving specific challenges encountered in abstractive summarization systems and enhancing the quality of the summaries produced.

1) Attention: The attention mechanism was initially inspired by the human visual attention system and has since become a fundamental component in neural network models, especially in natural language processing (NLP) tasks [35]. The attention mechanism in neural networks is a critical advancement that enhances the encoder-decoder architecture, particularly in tasks like abstractive text summarization. It allows the model to focus on different parts of the input sequence for each step of the output sequence, thereby capturing more nuanced relationships within the text. It addresses the limitation of traditional sequence-to-sequence models by enabling the network to weigh and focus on different parts of the input sequence, which is crucial for understanding long and complex texts. In the context of abstractive text summarization, attention mechanisms have been shown to significantly improve the quality of generated summaries. Krantz and Kalita [36] demonstrated the effectiveness of attention-based models in generating abstractive sentence summaries, highlighting the importance of this mechanism in capturing the essential elements of the source text. The attention mechanism can be mathematically represented as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2)$$

Here, Q, V, and K represent the query, value, and key matrices, respectively. The softmax function is applied to the scaled dot-product of Q and K, which determines the weightage of each part of the input sequence. The output is then computed as a weighted sum of the values V, where the weights are given by the attention scores. A representation of this procedure is illustrated in Fig. 8.

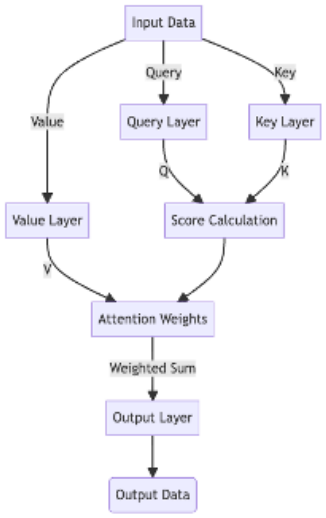

Fig. 8. Attention mechanism structure.

Fig. 8 illustrates how the attention mechanism operates within a neural network. It shows the flow of information from the input sequence through the attention module, where the query, key, and value matrices are computed and used to generate the attention scores. These scores are then applied to the input sequence to produce a context vector, which is used by the decoder to generate the output sequence. This mechanism is particularly effective in tasks like abstractive text summarization, where understanding the context and relationships within the text is crucial for generating coherent and accurate summaries.

a) Self-attention Mechanism: Self-attention, also known as intra-attention, is a mechanism that enables a model to assign varying degrees of importance to different segments of input data in relation to one another. This feature has been pivotal in developing architectures like the Transformer, which heavily utilizes self-attention for processing sequences [37]. In the realm of natural language processing (NLP), self-attention has enhanced the performance of tasks including machine translation, text summarization, and sentiment analysis. Networks employing self-attention are capable of linking words that are far apart through shorter paths within the network than those used by RNNs, potentially enhancing their performance in capturing long-distance relationships between elements in the data [38]. Yang et al. [39] applied self-attention to identify relationships between sentences and introduced a copying mechanism to address the issue of words that are out of vocabulary (OOV). Duan et al. [40] introduced a contrastive attention mechanism within the sequence-to-sequence framework for the task of abstractive sentence summarization, aimed at creating concise summaries of source sentences. This mechanism comprises two types of attention: the traditional attention, which focuses on the relevant parts of the source sentence, and an opposing

attention, which targets the irrelevant or less significant parts. These attentions are trained inversely to enhance the impact of traditional attention while reducing the influence of the opposing attention through an innovative use of softmax and softmin functions. Furthermore, Wang [41] developed a model for abstractive text summarization that integrates a hybrid attention mechanism, leveraging sentence-level attention to refine the distribution of word-level attention, thereby improving ROUGE scores and preserving critical information in the summaries.

2) Copying: The copying mechanism has been effectively utilized in various abstractive text summarization models to enhance their ability to generate accurate and contextually relevant summaries. This mechanism allows models to directly copy words or phrases from the source text into the summary, ensuring the inclusion of key information and terminology [24]. The copying mechanism in neural networks, particularly in sequence-to-sequence models, can be represented by a formula that combines the probabilities of generating words from the vocabulary and copying words from the source text. The formula typically used is:

$$P(w) = (1 - p_{\text{gen}}) \cdot \sum_{i:w_i=w} a_i^t \tag{3}$$

Here:

- $P(w)$ is the probability of the word w being in the output sequence.
- p_gen is the generation probability (used in the pointer generator mechanism, but here we focus on the copying part, hence $1 - p_gen$.
- a_i^t represents the attention distribution, where i indexes over the input sequence at time t.
- The summation $\sum_{i:w_i=w} a_i^t$ accumulates the attention scores for all instances of the word w in the input sequence, contributing to the probability of copying the word w from the input.

Li et al. [42] introduced the Correlational Copying Network (CoCoNet) for abstractive summarization. CoCoNet enhances the standard copying mechanism by tracking the copying history, thereby encouraging the model to copy input words relevant to previously copied ones. Zhou et al. [43] developed SeqCopyNet, a framework that not only learns to copy single words, but also copies sequences from the input sentence. This model leverages pointer networks to select sub-spans from the source text, integrating sequential copying into the generation process. These studies and applications demonstrate the effectiveness of the copying mechanism in enhancing the quality of abstractive text summarization. By allowing direct copying from the source text, these models can produce summaries that are both accurate and reflective of the original content.

3) Coverage: The coverage mechanism in neural networks, particularly in sequence-to-sequence models for tasks like abstractive text summarization, is designed to tackle the issue of repetition and improve the focus of the model on different parts of the input text [24]. This mechanism keeps track of what has been covered in the source text, thereby preventing the model from repeatedly attending to the same parts of the input. The coverage mechanism can be mathematically represented as follows:

$$c_t = \sum_{i=0}^{t-1} a_i \tag{4}$$

Here:

- c_t is the coverage vector at time step t, which accumulates the attention weights a_i from all previous time steps.
- a_i represents the attention distribution at time step i.
- The summation $\sum_{i=0}^{t-1} a_i$ accumulates the attention distributions from all previous time steps up to $t - 1$.

The coverage vector ct is then used to inform the attention mechanism at each decoding step, helping the model to distribute its attention more evenly across the entire input sequence. See et al. [24] incorporated the coverage mechanism into their pointer-generator network for abstractive summarization, significantly reducing the issue of repetition in the generated summaries. This approach demonstrated the effectiveness of the coverage mechanism in improving the quality and readability of machine-generated summaries.

4) Pointer-Generator: The Pointer-Generator model is designed to tackle the challenges associated with out-of-vocabulary (OOV) words and inaccuracies in reproducing factual information. It manages this by either replicating words or factual data through a pointer mechanism or by generating new words via a generator component [24]. The model calculates both an attention distribution and a vocabulary distribution (P_{vocab}), along with a generation probability denoted as p_gen. This generation probability determines whether the next word will be generated from the model's own vocabulary or copied directly from the source text. The process for determining the final probability of any given output word is established through a specific mathematical formulation:

$$P(w) = p_{\text{gen}} \cdot P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \cdot \sum_{i:w_i=w} a_i^t \tag{5}$$

Here:

- $P(w)$ is the probability of the word w being in the output sequence.
- P_gen is the generation probability, which is a scalar learned by the model. It decides whether to generate a word from the vocabulary or copy from the source text.
- $P_{\text{vocab}}(w)$ is the probability of the word w according to the model's vocabulary distribution.
- a_i^t represents the attention distribution, where i indexes over the input sequence. It indicates the model's focus on different parts of the input sequence at time t.
- The summation $\sum_{i:w_i=w} a_i^t$ accumulates the attention scores for all instances of the word w in the input sequence, contributing to the probability of copying the word w from the input.

Ren and Zhang [44] proposed a pointer-generator text summarization model that integrates part of speech features. This model uses a pointer-generator network to control whether to generate or copy words, effectively addressing out-of-vocabulary issues and avoiding duplication problems. Liu et al. [45] proposed a topic-aware architecture to adapt the Pointer-Generator model for summarizing conversations. Rehman et al. [26] used Pointer-Generator networks with SciBERT embeddings for automatic research highlight generation.

C. Training and Optimization

Training refers to the method through which a model acquires knowledge. Sequence-to-sequence models require training to predict the subsequent word in a sequence, based on the preceding output and the contextual information.

1) Word level training: Word level training in abstractive text summarization involves focusing on individual words, their representations, and relationships. This training method is crucial for understanding the semantics and syntactic properties of words within the context of summarization. Word level training is fundamental in developing models that can accurately interpret and reproduce the meaning of words in summaries. It often involves techniques like word embeddings to capture semantic relationships between words [46].

2) Sequence level training: Sequence level training in abstractive text summarization involves training models to understand and process sequences of words such as sentences or paragraphs. This type of training is crucial for models to capture the context and flow of ideas in the text. Unlike word level training, which focuses on individual words, sequence level training helps the model understand how words combine to form meaningful phrases and sentences. Sequence level training deals with sequences of words, such as sentences or paragraphs. It is essential for understanding the context and how words are used together in sequences. Such training is crucial for enabling models to understand the narrative structure and context present in the text, allowing them to produce summaries that are both coherent and relevant to the context [47].

3) Document level training: Document-level training involves training models on entire documents to understand the overall theme, structure, and sentiment. This approach is crucial for tasks like abstractive text summarization, where the model needs to grasp the main ideas and narrative structure of the entire text. Fecht et al. [48] examined the impact of sequential transfer learning on abstractive machine summarization using multilingual BERT and highlighted the effectiveness of transfer learning in improving the summarization of texts in languages.

4) Sentence level training: Sentence-level training in abstractive text summarization focuses on understanding and processing individual sentences within a document. This approach is crucial for models to capture the meaning, structure, and nuances of each sentence, which is essential for generating coherent and contextually accurate summaries. Chen et al. [47] proposed a novel extractive-generative model for text summarization using synthetic seq-2-seq pairs. The model demonstrates promise at the sentence level, indicating the potential of sentence-level training in generating sensible output for summarization tasks under resource constraints.

5) Transfer learning: Transfer learning is a technique in machine learning where a model designed for one specific task is repurposed as the foundation for a model on a different task. Within the realm of abstractive text summarization, this method involves adopting a model that has been pre-trained on a broad and varied dataset, and then fine-tuning this model for the specialized task of text summarization [29]. Zolotareva et al. [49] investigated the application of Sequence-to-sequence recurrent neural networks and Transfer Learning with the Unified Text-to-Text Transformer in abstractive text summarization, demonstrating significant enhancements in the summarization process.

6) Reinforcement learning: Reinforcement learning in abstractive text summarization is a training approach where the model learns to make decisions, such as selecting the most relevant content for the summary, by receiving feedback in the form of rewards or penalties. Nguyen et al. [8] explored a performance-driven reinforcement learning approach for abstractive text summarization, demonstrating its effectiveness in improving summary quality. Buciumas [50] discusses the use of reinforcement learning in abstractive text summarization, focusing on pre-trained models and RL to generate summaries across multiple datasets.

D. Datasets

Datasets play a crucial role in the training and evaluation of models, particularly in the field of abstractive text summarization. In the English language, there are several key datasets available for this purpose. The CNN/Daily Mail dataset, which includes articles and editorials from CNN and Daily Mail, was introduced for abstractive summarization by Nallapati et al. [46]. This dataset comprises 286,817 training pairs, 13,368 validation pairs, and 11,487 test pairs.

Another significant dataset is Newsroom, introduced by Grusky et al. [51], which contains 1.3 million articles and summaries authored by professionals from 38 different news publications, covering news from 1998 to 2017. The Gigaword dataset is an extensive collection of English newswire text. First used for abstractive summarization by Rush et al. [52], it includes approximately 9.5 million news articles. In Gigaword, each article's first sentence and its headline are used to form a source-summary pair. The Document Understanding Conference (DUC) dataset, another vital resource, is split into two parts: the 2003 corpus with 624 document-summary pairs and the 2004 corpus with 500 pairs, as noted by Nallapati et al. [46].

Zhang et al. [53] introduced MAC-SUM, a human-annotated summarization dataset for controlling mixed attributes. The XSUM dataset consists of BBC articles, each accompanied by a single-sentence summary [54]. It contains over 200,000 BBC articles each accompanied by a single-sentence summary. These summaries are professionally written, often serving as introductory or headline sentences. This dataset is designed specifically for the task of extreme summarization, a form of abstractive summarization that aims to produce a single-sentence summary for each document. MAC-SUM comprises source texts from two sectors namely news articles and dialogues, accompanied by human-generated summaries that are regulated according to five specific attributes: Length, Extractiveness, Specificity, Topic, and Speaker.

Table II provides a concise overview of the datasets used in abstractive text summarization, highlighting their origins, contents, and unique features. These datasets are instrumental in training and evaluating models in the field, offering diverse and comprehensive resources for developing advanced summarization techniques.

E. Evaluation Metrics

Automating the summarization task necessitates a system or method for its assessment and evaluation. While manual assessment is one approach, there are also specific metrics designed for this purpose. ROUGE (Recall Oriented Understudy for Gisting Evaluation), introduced by Lin [55], focuses on recall and is widely used for evaluating automatic summaries. ROUGE is primarily based on recall. It measures the overlap of n-grams, word sequences, and word pairs between the generated summary and a set of reference summaries. ROUGE is the most popular metric in summarization tasks due to its effectiveness in capturing content overlap. It is particularly useful for evaluating the extent to which the key information from the source text is retained in the summary. ROUGE's widespread adoption in summarization research is attributed to its alignment with human judgment, especially in terms of content coverage and informativeness.

BLEU (Bilingual Evaluation Understudy), proposed by Papineni [56], evaluates based on precision and recall, and its scores are commonly applied in automatic summarization system assessments. METEOR (Metric for Evaluation of Translation with Explicit Ordering), developed by Banerjee and Lavie [57], is primarily for assessing machine translation outputs but is also applicable to summarization. BLEU evaluates based on precision. It compares the n-grams of the generated summary with those in the reference summary and calculates a score based on the proportion of n-grams in the generated summary that appear in the reference summary. Although originally developed for machine translation, BLEU is also used in summarization. It is particularly effective for assessing the preciseness of the generated summaries in capturing the essential points of the source text.

METEOR utilizes modified precision and recall. These evaluation metrics offer an estimate of the extent to which the auto-generated summary aligns with the reference summary. Table III provides an overview of the key metrics used in the evaluation of automatic text summarization systems, highlighting their foundational principles and applications. METEOR is based on modified precision and recall. It goes beyond mere lexical matching to include stemming and synonymy, providing a more nuanced evaluation of translation outputs and summaries. While primarily designed for machine translation, METEOR's application in summarization is valuable, especially for its ability to recognize paraphrases and semantically equivalent phrases, thus offering a more comprehensive evaluation.

IV. Discussion

This section presents the analysis and findings of the survey, and addresses significant topics of discussion.

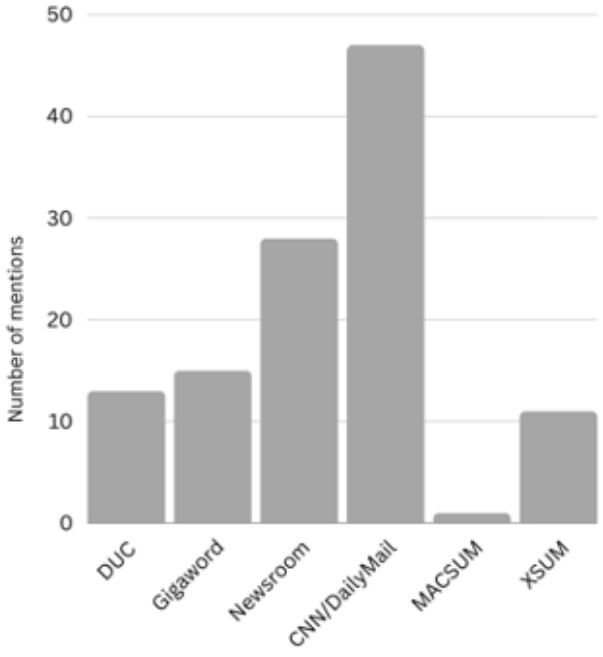

Fig. 9. Number of studies per dataset.

A. Dataset

As depicted in Fig. 9, it is clear that numerous studies in the survey predominantly use the CNN/Daily Mail dataset for abstractive summarization tasks. Further analysis of the dataset aspect reveals that this survey identifies several large datasets in languages deemed to be low-resource, such as Turkish [32], Hungarian [32], Indonesian [58], and French [59].

B. Evaluation Metrics

The evaluation of generated summaries is crucial for assessing the effectiveness and accuracy of different models and approaches. Several metrics have been developed for this purpose, each with its unique focus and methodology. Among these, ROUGE, BLEU, and METEOR are the most prominent. Notably, we observed among the papers that ROUGE is the most widely used metric in this domain as presented in Table IV. Our findings on the frequency of usage of evaluation metrics employed by researchers underscore the predominant preference for ROUGE as the primary evaluation metric, with 53 instances of use among the collected studies. This preference starkly contrasts with the 17 instances where BLEU (Bilingual Evaluation Understudy) was utilized and the three instances of METEOR (Metric for Evaluation of Translation with Explicit Ordering) application.

The preference for ROUGE as the dominant metric for evaluating generated summaries within ATS research can be attributed to several factors. Firstly, ROUGE's design specifically caters to summary evaluation by measuring the overlap of n-grams between the generated summaries and reference summaries. This characteristic makes ROUGE highly suitable for tasks where capturing the gist of the text is more critical than the exact reproduction of phrases or sentence structures, aligning well with the objectives of ATS. Secondly, ROUGE

TABLE II. KEY DATASETS FOR ABSTRACTIVE TEXT SUMMARIZATION RESEARCH

Dataset Name	Description	Authors	Details
CNN/Daily Mail	Articles and editorials from CNN and Daily Mail for abstractive summarization.	Nallapati et al. [46]	Comprises 286,817 training pairs, 13,368 validation pairs, and 11,487 test pairs.
Newsroom	1.3 million articles and summaries from 38 news publications.	Grusky et al. [51]	Covers news from 1998 to 2017.
Gigaword	Extensive collection of English newswire text.	Rush et al. [52]	Approximately 9.5 million news articles. Source-summary pairs created from the first sentence and headline.
DUC	Document Understanding Conference dataset for text summarization.	Nallapati et al. [46]	Two parts: 2003 corpus with 624 pairs and 2004 corpus with 500 pairs.
MAC-SUM	Human-annotated summarization dataset for controlling mixed attributes.	Zhang et al. [53]	Contains texts from news articles and dialogues with summaries controlled by attributes.
XSUM	BBC articles each accompanied by a single-sentence summary for extreme summarization.	Narayan et al. [54]	Over 200,000 articles, each with a professionally written single-sentence summary.

TABLE III. EVALUATION METRICS FOR ABSTRACTIVE TEXT SUMMARIZATION

Metric	Description	Year	Application
ROUGE	Based on recall, evaluates the quality of automatic summaries.	2004	Commonly used in automatic summarization evaluation.
BLEU	Evaluates based on precision and recall, used for summarization and translation.	2002	Applied in automatic summarization system assessments.
METEOR	Utilizes modified precision and recall, originally for machine translation but applicable to summarization.	2005	Used for both machine translation and summarization evaluation.

offers various measures such as ROUGE-N (for n-gram overlap), ROUGE-L (for the longest common subsequence), and ROUGE-W (for weighted longest common subsequence), providing a comprehensive assessment framework that can accommodate different summarization goals. This versatility ensures a broader evaluation perspective, covering aspects from simple word overlap to more complex semantic coherence and fluency. In contrast, BLEU, while widely used in machine translation evaluation, focuses more on precision—the proportion of words in the generated text that appear in the reference texts. This metric's emphasis on precision over recall can be less aligned with the summarization task's nuances, where capturing the most relevant information (regardless of the exact wording) is often more critical.

METEOR, despite offering a more balanced evaluation by considering synonymy and stemming and aiming for higher correlation with human judgment, is used less frequently. This lesser usage might stem from its complexity and computational demand, making ROUGE a more straightforward and efficient choice for many researchers. These SLR results indicate a clear consensus within the ATS research community on the effectiveness and appropriateness of ROUGE for evaluating summarization tasks. The findings also suggest a need for continuous evaluation of existing metrics and the development of new metrics that can capture the qualitative aspects of summaries more accurately, reflecting human judgments and preferences.

TABLE IV. FREQUENCY OF USAGE OF EVALUATION METRICS AMONG COLLECTED STUDIES

Metric	Number of Usages
ROUGE	53
BLEU	17
METEOR	3

C. Model Performance Comparison

Some recent abstractive text summarization approaches with highest Rouge scores were selected and compared as seen in Table V. Among the selected studies, Zhao et al. [60] which focused on sequence likelihood calibration achieved the highest ROUGE scores across all three metrics: ROUGE-1 (48.88), ROUGE-2 (24.94), and ROUGE-L (45.76). ROUGE-1 measures the overlap of unigrams between the generated summaries and the reference summaries, indicating the accuracy of capturing key points. ROUGE-2 evaluates the overlap of bigrams, reflecting the model's ability to preserve essential phrases and the coherence of the generated text. Lastly, ROUGE-L assesses the longest common subsequence, focusing on the fluency and the structural similarity between the generated summaries and the reference summary. This indicates the superior ability of their model to align closely with reference summaries, particularly in terms of content overlap and fluency. Their approach demonstrated the effectiveness of integrating contrastive learning into summarization tasks. He et al. [61] address computational efficiency, a critical aspect in processing long sequences. Their innovative use of the Fast Fourier Transform (FFT) operator showcases how computational advancements can enhance model performance. Wang et al. [62] introduced a concept of salience allocation as guidance, highlighting the importance of content selection in generating coherent summaries. Lastly, Ravaut et al. [63] with "SummaReranker" introduced a re-ranking framework, conducting summary generation by selecting better candidates from a set of options. This approach underscores the potential of multi-stage processing in improving summary quality.

A notable trend is the focus on enhancing existing models through innovative techniques like contrastive learning, sequence likelihood calibration, and re-ranking strategies. These methods aim to refine the model's ability to generate summaries that are not only accurate, but also contextually rich and coherent. The studies also reflect a growing interest in addressing specific challenges such as computational efficiency and the quality of content selection, which are crucial for the practical application of summarization models. The analysis of these studies reveals a dynamic and rapidly evolving field, with each approach contributing to the overall goal of improving the quality and applicability of abstractive text summarization

TABLE V. ROUGE SCORES FOR VARIOUS APPROACHES

Authors	Approach	ROUGE-1	ROUGE-2	ROUGE-L
Liu and Liu [45]	SimCLS: Contrastive Learning Framework	46.67	22.15	43.54
He et al. [61]	Fourier Transformer: Removing Sequence Redundancy	44.76	21.55	41.34
Wang et al. [62]	SEASON: Salience Allocation Guidance	46.27	22.64	43.08
Ravaut et al. [63]	SummaReranker: Re-ranking Framework	47.16	22.61	43.87
Zhao et al. [60]	SLiC: Sequence Likelihood Calibration	48.88	24.94	45.76
Liu et al. [52]	BRIO: Non-Deterministic Distribution Training	47.78	23.55	44.57

models. The diversity in methodologies and the continuous push for higher ROUGE scores indicate a vibrant research landscape, driven by the pursuit of models that can generate summaries with high fidelity to the original content and contextual relevance.

To provide a more comprehensive view, we expanded the comparison by evaluating the selected models on multiple benchmark datasets, including CNN/DailyMail, XSum, and Newsroom (Table VI). Each dataset poses unique challenges and helps illustrate the practical implications of the advancements in summarization techniques. The detailed ROUGE scores presented in Table VI offer a comprehensive comparison of various abstractive text summarization approaches across multiple datasets, including CNN/DailyMail, XSum, and Newsroom. Zhao et al. [60], utilizing Sequence Likelihood Calibration (SLiC), demonstrate superior performance with the highest ROUGE-1, ROUGE-2, and ROUGE-L scores across both the CNN/DailyMail and XSum datasets, indicating their model's robust generalization and efficacy in different contexts. Liu and Liu [45] with SimCLS also show strong performance, particularly on the XSum dataset, reflecting the effectiveness of contrastive learning frameworks. He et al. [61]'s Fourier Transformer approach, while efficient, lacks data for the XSum and Newsroom datasets, highlighting a gap in evaluation. Wang et al. [62] and their SEASON model excel in the Newsroom dataset, particularly in ROUGE-2, suggesting a strength in handling diverse and complex data. Ravaut et al. [63]'s SummaReranker performs well across the board, especially on XSum, emphasizing the potential of re-ranking strategies. Liu et al. [52] with BRIO maintain competitive scores, showcasing consistent performance across multiple datasets. This diverse range of approaches and their respective performances underscore the evolving landscape of abstractive summarization, where each model brings unique strengths to address the multifaceted challenges of summarizing varied content types.

The practical implications of these advancements are significant. Zhao et al. [60] and their sequence likelihood calibration technique consistently achieve high ROUGE scores across various datasets, indicating a robust approach that generalizes well. However, their method may involve higher computational costs due to the complexity of the calibration process. He et al. [61] emphasize computational efficiency, which is particularly advantageous for real-time applications and processing long documents. Their use of the Fast Fourier Transform (FFT) operator reduces the computational burden, making it a practical choice for scenarios where efficiency is critical. Wang et al. [62] focus on content selection through salience allocation, which enhances the relevance and coherence of the summaries. This method is especially useful for summarizing documents where specific information needs to be prioritized. Ravaut et al. [63] and their re-ranking framework show the potential of multi-stage processing in improving summary quality. By selecting the best candidates from a set of generated summaries, their approach can produce more refined and accurate summaries. Liu and Liu [45] with their contrastive learning framework and Liu et al. [52] with their non-deterministic distribution training approach also contribute to the field by exploring different aspects of model training and summary generation, offering diverse solutions to common challenges. The detailed comparison across multiple datasets and the discussion of practical implications provide a clearer picture of the strengths and weaknesses of each approach. This helps in understanding the trade-offs involved and guides the selection of appropriate models for specific summarization tasks.

D. Emerging Issues and Challenges

While there have been notable advancements in abstractive text summarization using neural networks in recent times, these systems continue to present various challenges and issues for researchers. Gaining an understanding of these current problems and devising solutions to address them will lead to the development of more effective and dependable summarization systems. Based on reviewed literature, the following are key issues and challenges in abstractive text summarization:

1) Complexity of transformer-based models: Models like Transformers, while effective, suffer from quadratic complexity with respect to input text length. This makes processing long documents computationally expensive and less efficient [64]. While self-attention offers many benefits, such as the ability to capture long-range dependencies and parallelize computation, it also presents challenges. One significant issue is the quadratic computational cost relative to the sequence length, which can make it resource-intensive for very long sequences [37].

Researchers have been working on developing more efficient self-attention mechanisms to mitigate this issue. Bonnaerens and Dambre's [65] approach to addressing this challenge is the introduction of Learned Thresholds Token Merging and Pruning (LTMP), which combines token merging and pruning to reduce the number of input tokens that need processing, effectively lowering the computational load. This method leverages dynamic thresholding to determine the tokens to be merged or pruned, demonstrating significant efficiency improvements [65]. Additionally, Tang et al. [38] introduced QuadTree Attention mechanism, which presents a novel solution by reducing computational complexity from quadratic to linear. By constructing token pyramids and computing attention in a coarse-to-fine manner, QuadTree Attention focuses on relevant regions by selecting the top patches with the highest attention scores, thereby streamlining the attention process.

TABLE VI. DETAILED ROUGE SCORES ON MULTIPLE DATASETS

Authors	Approach	CNN/DailyMail			XSum			Newsroom		
		R1	R2	RL	R1	R2	RL	R1	R2	RL
Liu and Liu [45]	SimCLS: Contrastive Learning Framework	46.67	22.15	43.54	47.61	24.57	39.44	-	-	-
He et al. [61]	Fourier Transformer: Removing Sequence Redundancy	44.76	21.55	41.34	-	-	-	-	-	-
Wang et al. [62]	SEASON: Salience Allocation Guidance	46.27	22.64	43.08	-	-	-	46.00	33.37	42.03
Ravaut et al. [63]	SummaReranker: Re-ranking Framework	47.16	22.55	43.87	48.12	24.95	40.00	-	-	-
Zhao et al. [60]	SLiC: Sequence Likelihood Calibration	47.97	24.18	44.88	49.77	27.09	42.08	-	-	-
Liu et al. [52]	BRIO: Non-Deterministic Distribution Training	47.78	23.55	44.57	49.07	25.59	40.40	-	-	-

Wu et al. [66] introduced Singularformer, a transformative approach by leveraging neural networks to learn the singular value decomposition process of the attention matrix. This process aims to design a linear-complexity and memory-efficient global self-attention mechanism, demonstrating favorable performance against other Transformer variants with lower time and space complexity. Continued exploration and innovation in model optimization, attention mechanism refinement, and hardware acceleration are critical for advancing the field and expanding the applicability of these powerful models.

2) Model hallucination: Model hallucination represents a formidable challenge in abstractive text summarization, where models often generate text that deviates factually from the input, undermining the reliability and accuracy of the summaries. This issue not only questions the credibility of automated summarization but also poses significant hurdles in applications requiring high factual consistency [64]. Recent research has introduced innovative approaches to mitigate this problem. Contrastive Parameter Ensembling (CaPE) offers a promising solution by leveraging variations in training data noise. By fine-tuning a base model on subsets of clean and noisy data, CaPE effectively reduces hallucination, enhancing factual accuracy across different datasets [14]. Similarly, another study proposes training augmentation methods for image captioning to reduce object bias, a form of hallucination, without increasing model size or requiring additional training data [7]. Further, a simple yet effective strategy proposed for neural surface realization addresses content hallucination by integrating language understanding modules for data refinement, significantly reducing unaligned noise and improving content correctness [67]. Additionally, the Chain of Natural Language Inference (CoNLI) framework has been developed for detecting and mitigating ungrounded hallucinations in large language models, showcasing an effective method for enhancing text quality through rewrite [68].

These advancements highlight the community's ongoing efforts in confronting and reducing model hallucination in text generation tasks. By focusing on data quality, leveraging contrastive learning, and integrating understanding mechanisms, researchers continue to push the boundaries of what is possible in generating accurate and reliable automated summaries.

3) Domain shift: The performance of models often degrades when the distribution of the training and test corpus is not the same. This domain shift is particularly problematic in domain-specific summarization tasks [64]. To address the challenge of domain shift, researchers have been exploring a variety of techniques aimed at enhancing the adaptability of models across different domains. One promising approach is the leveraging of pretrained transformer models, which has shown remarkable versatility in various NLP tasks. For instance, Zhang et al. [19] demonstrates the potential of fine-tuning BART on domain-specific datasets to generate fluent and adequate summaries of doctor-patient conversations. Their methodology effectively overcomes the obstacles posed by domain shift, limited training data, and the inherent variability of target summaries. By integrating these strategies, researchers are making strides towards developing models that maintain high performance levels across varying domains, thus expanding the applicability and reliability of automatic text summarization technologies.

4) Quality of datasets: The effectiveness of summarization models is closely tied to the quality of datasets they are trained on. Srivastava et al. [69] highlighted issues like information coverage, entity hallucination, and the inherent complexity of summarization tasks as significant challenges that can impact model performance. These issues underscore the need for high-quality datasets that accurately represent the diversity and complexity of real-world texts and the necessity for summarization models to generate accurate, reliable, and coherent summaries. Information coverage is essential for ensuring that all relevant aspects of the source document are represented in the summary. This necessitates datasets that are comprehensive and reflective of the variety of information that summaries should convey. To improve information coverage, Utama et al. [70] developed Falsesum, a data generation pipeline that introduces factual inconsistencies in summaries to train models to better recognize and avoid such errors.

In exploring the balance between lexical and semantic quality in summarization, Sul and Choi [71] proposed a training method incorporating a re-ranking system. This approach aims to mitigate false positives in ranking, enhancing the model's ability to interpret the meaning of summaries without compromising lexical quality. Moreover, Liu et al. [52] introduced a training paradigm that assumes a non-deterministic distribution for candidate summaries. By assigning probability mass based on quality, this method aims to order abstractive summarization more effectively, showcasing an innovative approach to handling the complexity of summarization tasks. The complexity of summarization tasks, with varying degrees of abstraction, summarization length, and domain specificity, calls for datasets that capture this diversity. Adams et al. [72] explored the characteristics of effective calibration sets in training, finding that certain strategies, like maximizing metric margins and minimizing surprise, can improve model

performance across different summarization tasks. To address these challenges, research directions included the development of advanced techniques for generating high-quality, diverse datasets and training models that are more adept at handling the intricacies of summarization.

5) Factual inconsistency: The abstraction ability of neural models can lead to the distortion or fabrication of factual information, causing inconsistency between the original text and the summary. This issue necessitates the development of fact-aware evaluation metrics and summarization systems [73]. The abstraction capabilities of neural models, while enabling the generation of concise and coherent summaries, often lead to the distortion or fabrication of factual information. This misalignment between the original text and the generated summary, known as factual inconsistency, undermines the reliability of summarization systems. Huang et al. [73] underscore the necessity for the development of fact-aware evaluation metrics and systems that can ensure the factual accuracy of summaries. To mitigate these challenges, researchers have been exploring various methodologies.

Li and Xu [59] proposed a clinical trial prediction-based factual inconsistency detection approach tailored for medical text summarization. This novel methodology leverages the relationship between clinical trial outcomes and the factual consistency of related medical articles' summaries. By predicting the success or failure of clinical trials based on summaries, their approach offers a direct method to assess factual consistency, showcasing a specialized application of factual accuracy evaluation in the medical domain. Utama et al. [70] introduced Falsesum, a data generation pipeline that creates document-level natural language inference (NLI) examples specifically designed to recognize factual inconsistencies in summarization. This approach enhances models' ability to discern and avoid factual errors by incorporating high-quality, task-oriented examples into their training data, addressing the need for datasets that challenge models to maintain factual integrity.

Exploring the capabilities of large language models, Luo et al. [74] investigated ChatGPT's potential as a factual inconsistency evaluator for abstractive text summarization. Their findings suggest that ChatGPT, under a zero-shot setting, can outperform state-of-the-art evaluation metrics across various factuality evaluation tasks. This highlights the promise of leveraging advanced language models for more nuanced and effective factual consistency assessments in summarization. To further mitigate factual inconsistency, future research directions may involve the integration of fact-checking modules within summarization frameworks, the development of more advanced fact-aware training methodologies, and the exploration of novel dataset augmentation techniques. These strategies aim to refine the summarization process, ensuring that models can produce summaries that are not only coherent and concise, but also factually accurate.

6) Multimodal summarization: Integrating multimodal knowledge, such as combining text and images for abstractive text summarization, represents a significant advancement in the field, yet it is fraught with challenges. The primary difficulty lies in bridging the semantic gaps between different modalities, which can hinder the effective fusion of multimodal data. This issue is particularly pronounced due to the divergent nature of information conveyed through text and images, necessitating innovative approaches to achieve a coherent and comprehensive summary that leverages both modalities effectively [6].

To tackle these challenges, Liang et al. [75] introduced the D2TV framework. This innovative approach, aimed at Many-to-Many Multimodal Summarization (M3S), leverages dual knowledge distillation and target-oriented vision modeling to enhance both multimodal monolingual summarization (MMS) and multimodal cross-lingual summarization (MXLS) tasks. Their framework demonstrates the effectiveness of mutual knowledge transfer between MMS and MXLS, alongside employing a contrastive objective to refine visual features for summarization, showcasing a promising direction in multimodal summarization research.

He et al. [61] developed the A2Summ model, which introduces a unified approach to align and attend to multimodal inputs. Their work focuses on leveraging dual contrastive losses to model the correlations within and between samples across modalities, thereby enhancing the quality of multimodal summaries. This method highlights the importance of understanding and aligning multimodal information to generate reliable and high-quality summaries. These efforts represent a concerted move towards overcoming the inherent challenges of multimodal summarization. By developing models that can effectively process and integrate information from diverse modalities, researchers aim to generate summaries that are not only more informative and comprehensive, but also more engaging for users.

7) Low-Resourced languages: Abstractive text summarization for low-resourced languages like Urdu faces challenges due to the lack of extensive research and datasets. Generating abstractive summaries in such languages demands more focused research efforts [76]. Abstractive text summarization in low-resourced languages, such as Urdu, presents distinct challenges due to the scarcity of extensive research, datasets, and computational resources tailored to these languages. The development of abstractive summarization capabilities in such contexts is hindered by the lack of high-quality, large-scale datasets, and advanced NLP tools that are readily available for languages with more substantial digital footprints [76]. This gap in resources and research attention limits the ability to apply state-of-the-art NLP methodologies, including deep learning techniques, which have shown significant success in abstractive summarization tasks in languages like English.

Baykara and Güngör [32] addressed this gap by introducing new large-scale datasets for agglutinative languages like Turkish and Hungarian, showcasing the potential for enhancing abstractive text summarization in languages that have traditionally been underrepresented in NLP research. Shafiq et al. [76] delved into the challenges and solutions for abstractive text summarization of Urdu using deep learning, highlighting the need for dedicated efforts to improve summarization techniques for low-resourced languages. Mascarell et al. [77] proposed entropy-based sampling approaches for abstractive multi-document summarization in low-resource settings, demonstrating innovative methods to address the challenges of summarizing content in languages with limited datasets. Hasan et al. [78] contributed to this field with XL-Sum, a large-scale multilingual abstractive summarization dataset covering 44 languages, many of which are low-resourced. This

initiative marks a significant step towards fostering research and development in multilingual summarization. Rodzman et al. [85] explored the use of text summarization as a positive hierarchical fuzzy logic ranking indicator for domain-specific retrieval of Malay translated Hadith, illustrating the application of summarization techniques in religious text analysis.

In the realm of Arabic text summarization, several notable approaches have been proposed to address the unique challenges posed by the Arabic language. Abdelwahab et al. [83] focused on using pre-processing methodologies and techniques to enhance Arabic text summarization. Their work emphasizes the importance of tailored pre-processing steps to handle the morphological and syntactical complexities of Arabic. Fejer and Omar [86] introduced a method combining clustering and keyphrase extraction for automatic Arabic text summarization. This approach leverages clustering to group similar text segments and keyphrase extraction to identify the most salient information, thereby improving the coherence and relevance of the summaries. These efforts reflect a growing interest in developing robust summarization techniques for Arabic, addressing the linguistic challenges and contributing to the broader field of natural language processing for underrepresented languages. These studies underscore the growing interest and ongoing efforts to extend abstractive summarization capabilities to low-resourced languages, aiming to close the gap in NLP research and application across linguistic landscapes.

8) Evaluation metrics: The current evaluation metrics, such as ROUGE, while widely used, may not fully encapsulate the quality of generated summaries, especially in capturing nuances that align with human judgment. Srivastava et al. [69] underlined the necessity for more comprehensive and nuanced evaluation methods that can better reflect the quality perceived by humans. This calls for the development of metrics that go beyond traditional approaches to evaluate the effectiveness of summarization systems in producing summaries that are not only relevant, but also coherent and faithful to the original text. Dash et al. [79] proposed evaluating summarization algorithms from a new perspective that considers fairness in representation across different socially salient groups. Their work introduces the novel fairness-preserving summarization algorithm 'FairSumm', which aims to produce high-quality summaries while ensuring equitable representation, marking a step beyond traditional ROUGE-centric evaluations. Gao et al. [80] proposed SUPERT, an unsupervised evaluation metric for multi-document summarization that rates summary quality by measuring its semantic similarity with a pseudo reference summary. SUPERT's use of contextualized embeddings and soft token alignment techniques represents a move towards more semantically rich evaluation frameworks.

The development of these comprehensive evaluation frameworks is crucial for advancing the field of text summarization and ensuring that generated summaries meet high standards of quality and utility. The exploration of evaluation metrics and methods that better capture the quality of generated summaries as perceived by humans is crucial for advancing the field of text summarization. As summarization systems become increasingly sophisticated, the development of equally sophisticated evaluation metrics will be essential to ensure their effectiveness and utility.

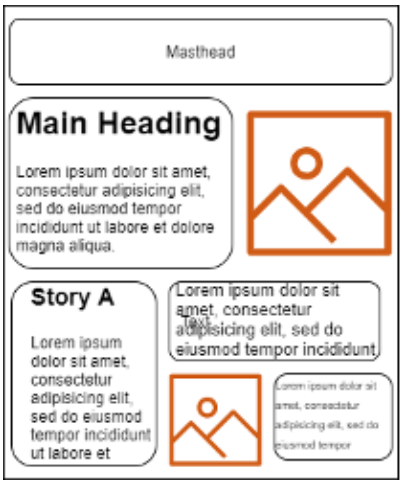

Fig. 10. Newspaper summary slot example.

9) Summary length control: Controlling the length of the generated summaries is a significant challenge in Abstractive text summarization. Models often struggle to produce summaries of a desired length while maintaining the essence and coherence of the original text. This issue is crucial for applications where space is limited or a specific summary length is required. Developing methods to effectively manage summary length without compromising content quality and relevance remains an area needing further exploration and innovation [81]. Despite recent solutions to control the length of the summary, this study observed that there is still an issue of arbitrarily doing so. While length embeddings can determine when to cease decoding, they do not specify which details ought to be encapsulated within the summary, given the length restriction [81]. Length embeddings merely incorporate length information on the decoder side, potentially overlooking crucial content because they fail to consider the elements that should be summarized under certain length limitations.

Previous studies have managed to control the length of summaries, setting it either as predefined [53] or flexible [7,81]. Although these approaches have improved the quality of length-constrained summarization, they all necessitate specifying a target length prior to generating the summary. In Saito et al. [81], the length of the prototype text must be determined before feeding it into their encoder-decoder model for generating a summary. Similarly, in the work of Takase and Okazaki [7], the remaining length must be specified at each step of the decoder in their Transformer-based encoder-decoder model. For instance, as illustrated in Fig. 10, when there is a need to summarize a lengthy newspaper story to fit a specific section on the newspaper cover, the existing methods would fall short as they rely on a predefined number of words for the summary length.

Predefined or arbitrary summary lengths present challenges, particularly when summaries need to fit specific spaces, such as a designated section on a magazine or newspaper cover. Current advanced models lack the capability to adapt summaries based on the size of the output area. For instance, these models struggle to condense a lengthy newspaper story into a brief summary that would precisely fill a specific part

of the newspaper cover, as their design relies on a fixed word count for summaries. This limitation is evident in the works of various researchers, including Zhao et al. [60], who have not addressed the need for space-constrained summarization.

Fan et al. [2] introduced the use of length embeddings at the start of the decoder to control summary length, offering a neural summarization model that allows for high-level attribute specification to tailor summaries more closely to user needs. Zhang et al. [19] developed a convolutional seq2seq model for summarization, enhancing the CNN with gated linear units (GLU), residual connections, and a hierarchical attention mechanism for simultaneous keyword and key sentence generation, alongside a copying mechanism for out-of-vocabulary (OOV) words. Takase and Okazaki [7] utilized positional encoding to indicate the remaining length at each step in their Transformer-based model.

Saito et al. [81] combined extractive and abstractive summarization by embedding an extractive model within an abstractive framework. This approach involves extracting a sequence of significant words ("prototype text") from the source, which then informs the summary generation process in the encoder-decoder model. Liu et al. [53] proposed a length-aware attention mechanism (LAAM) that tailors the encoding of the source text to the desired summary length, proving effective in creating high-quality summaries of various lengths, including lengths not encountered during training. Fein and Cuevas [82] proposed ExtraPhraseRank which used TextRank for sentence extraction and back-translation for word diversity, aiming to generate synthetic summaries with controlled lengths. While their approach shows modest improvements in ROUGE scores, the study also highlighted challenges in length control and the need for fine-tuning with human-written summaries. The aforementioned studies were able to provide a solution to the length control task; however, they all lacked the ability to self-determine the required summary length.

V. Conclusion

This study on abstractive text summarization has provided a comprehensive overview of the current state and advancements in the field. Through an in-depth analysis of various studies, we have identified key areas of focus, challenges, and innovative solutions that are shaping the future of abstractive text summarization. A framework for research was established, adhering to essential abstractive text summarization model design components such as encoder-decoder architecture, attention mechanisms, training and optimization methods, datasets, and evaluation metric. This framework is utilized to analyse abstractive summarization models, employing a concept matrix that underscores prevalent design trends in contemporary abstractive summarization systems.

The review highlighted significant progress in developing sophisticated models like Transformers and their variants, which have pushed the boundaries of abstractive text summarization. Studies have introduced novel approaches such as contrastive learning, sequence redundancy removal, and salience allocation, each contributing uniquely to the enhancement of summary quality.

The prevalent use of ROUGE as an evaluation metric was evident, with studies consistently aiming to improve ROUGE scores. However, the review also underscored the need for more nuanced evaluation metrics that can better capture the quality of generated summaries in terms of factual consistency, coherence, and alignment with human judgment. Despite remarkable progress, the field faces several challenges. These include model hallucination, domain shift, dataset quality, factual inconsistency, and the need for multimodal summarization. Particularly, the issue of summary length control emerged as a significant area needing further research, with various studies proposing different methods to address this challenge.

The review suggests that future research in abstractive text summarization should focus on developing more efficient models capable of handling long sequences, improving the factual accuracy of summaries, and creating better datasets, especially for low-resourced languages. Additionally, there is a clear need for more comprehensive evaluation frameworks that go beyond traditional metrics like ROUGE. The insights gained from this SLR provide a foundation for future research endeavors, aiming to overcome existing challenges and unlock new possibilities in the realm of automated text summarization.

References

[1] A. W. Palliyali, M. A. Al-Khalifa, S. Farooq, J. Abinahed, A. Al-Ansari, and A. Jaoua, "Comparative Study of Extractive Text Summarization Techniques," in *2021 IEEE/ACS 18TH INTERNATIONAL CONFERENCE ON COMPUTER SYSTEMS AND APPLICATIONS (AICCSA)*, 2021. doi: 10.1109/AICCSA53542.2021.9686867.

[2] A. Fan, D. Grangier, and M. Auli, "Controllable Abstractive Summarization," in *NEURAL MACHINE TRANSLATION AND GENERATION*, 2018, pp. 45–54.

[3] X. Wan, C. Li, R. Wang, D. Xiao, and C. Shi, "Abstractive Document Summarization via Bidirectional Decoder," in G. Gan, B. Li, X. Li, and S. Wang (Eds.), *ADVANCED DATA MINING AND APPLICATIONS, ADMA*, 2018.

[4] Q. Wang and J. Ren, "Summary-aware attention for social media short text abstractive summarization," *Neurocomputing*, vol. 425, pp. 290–299, 2021. doi: 10.1016/j.neucom.2020.04.136.

[5] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive Text Summarization: Enhancing Sequence-to-Sequence Models Using Word Sense Disambiguation and Semantic Content Generalization," *Computational Linguistics*, vol. 47, no. 4, pp. 813–859, 2021. doi: 10.1162/COLI_a_00417.

[6] Z. Zhang, C. Shu, Y. Chen, J. Xiao, Q. Zhang, and L. Zheng, "ICAF: Iterative Contrastive Alignment Framework for Multimodal Abstractive Summarization," 2021. doi: 10.1109/IJCNN55064.2022.9892884.

[7] S. Takase and N. Okazaki, "Positional Encoding to Control Output Sequence Length," 2019. Available: http://arxiv.org/abs/1904.07418.

[8] T.-P. N. Nguyen, N.-C. Van, and N.-T. Tran, "Performance-Driven Reinforcement Learning Approach for Abstractive Text Summarization," in *Advances in Intelligent Systems and Computing*, 2021.

[9] I. S. Blekanov, N. Tarasov, and S. S. Bodrunova, "Transformer-Based Abstractive Summarization for Reddit and Twitter: Single Posts vs. Comment Pools in Three Languages," *Future Internet*, vol. 14, no. 3, p. 69, 2022. doi: 10.3390/fi14030069.

[10] A. A. Syed, F. L. Gaol, and T. Matsuo, "A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization," *IEEE ACCESS*, vol. 9, pp. 13248–13265, 2021. doi: 10.1109/ACCESS.2021.3052783.

[11] N. Nazari and M. A. Mahdavi, "A survey on Automatic Text Summarization," *Journal of AI and Data Mining*, vol. 7, no. 1, pp. 121–135, 2019. doi: 10.22044/JADM.2018.6139.1726.

[12] N. Nazari and M. A. Mahdavi, "Specific Summarization Techniques," 2018.

[13] M. Zhang, G. Zhou, W. Yu, and W. Liu, "A Survey of Automatic Text Summarization Technology Based on Deep Learning," in

2020 INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND COMPUTER (ICAICE 2020), 2020, pp. 211–217. doi: 10.1109/ICAICE51518.2020.00047.

[14] Rahul, S. Rauniyar, and Monika, "A Survey on Deep Learning based Various Methods Analysis of Text Summarization," in *PROCEEDINGS OF THE 5TH INTERNATIONAL CONFERENCE ON INVENTIVE COMPUTATION (ICICT-2020)*, 2020, pp. 113–116.

[15] H. A. Shaffril Mohamed, S. F. Samsuddin, and A. Abu Samah, "The ABC of systematic literature review: the basic methodological guidance for beginners," *Quality and Quantity*, vol. 55, no. 4, pp. 1319–1346, 2021. doi: 10.1007/s11135-020-01059-6.

[16] J. A. Smith, L. R. Doe, and M. Q. Anderson, "Enhancing the rigor of research evaluations through comprehensive critical appraisal criteria," *Journal of Research Evaluation*, vol. 29, no. 4, pp. 475-488, 2020.

[17] Q. Ismail, K. Alissa, and R.M. Duwairi, "Arabic News Summarization based on T5 Transformer Approach," in *2023 14th International Conference on Information and Communication Systems (ICICS)*, pp. 1-7, 2023.

[18] A.R. Lubis, H.R. Safitri, Irvan, M. Lubis, M.L. Hamzah, A. Al-Khowarizmi, and O. Nugroho, "Enhancing Text Summarization with a T5 Model and Bayesian Optimization," *Revue d'Intelligence Artificielle*, 2023.

[19] Y. Zhang, D. Li, Y. Wang, Y. Fang, and W.D. Xiao, "Abstract Text Summarization with a Convolutional Seq2seq Model," *Applied Sciences*, 2019.

[20] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *Physica D: Nonlinear Phenomena*, 2018. doi: 10.1016/j.physd.2019.132306.

[21] H. A. Bouarara, "Recurrent Neural Network (RNN) to Analyse Mental Behaviour in Social Media," *International Journal of Social Science and Computational Intelligence*, 2021. doi: 10.4018/IJSSCI.2021070101.

[22] I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. E. Chapman, T. J. Amrhein, D. Mong, D. L. Rubin, O. Farri, and M. P. Lungren, "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," *Artificial Intelligence in Medicine*, 2019. doi: 10.1016/j.artmed.2018.11.004.

[23] J. Kumar, J. Mukherjee, R. Goomer, and A. K. Singh, "Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters," *Procedia Computer Science*, 2018. doi: 10.1016/J.PROCS.2017.12.087.

[24] A. See, P. J. Liu, and C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," 2017. Available: https://arxiv.org/abs/1704.04368.

[25] K. Cho, B. Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *SSST@EMNLP*, 2014.

[26] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, and P. Das, "Generation of Highlights From Research Papers Using Pointer-Generator Networks and SciBERT Embeddings," *IEEE Access*, 2023. doi: 10.1109/ACCESS.2023.3292300.

[27] S. Preethi, M.S. Krithick Shibi, S. Sheshan, R. Kingsy Grace, and M. Sri Geetha, "Abstractive Summarizer using Bi-LSTM," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, pp. 1605-1609, 2022.

[28] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Neural Information Processing Systems*, 2017.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, 2020.

[30] Q.A. Itsnaini, M. Hayaty, A.D. Putra, and N.A. Jabari, "Abstractive Text Summarization using Pre-Trained Language Model "Text-to-Text Transfer Transformer (T5)", *ILKOM Jurnal Ilmiah*, 2023.

[31] N. Darapaneni, R. Prajeesh, P. Dutta, V. K. Pillai, A. Karak, and A. Paduri, "Abstractive Text Summarization Using BERT and GPT-2 Models," in *2023 IEEE International Conference on Soft Computing and Pattern Recognition (ICSOFT)*, pp. 1605-1609, 2023. doi: 10.1109/IConSCEPT57958.2023.10170093.

[32] B. Baykara and T. Güngör, "Turkish abstractive text summarization using pretrained sequence-to-sequence models," *NATURAL LANGUAGE ENGINEERING*, vol. 29, no. 5, pp. 1275–1304, 2022. doi: 10.1017/S1351324922000195.

[33] V. Kieuvongngam, S. Wong, and C. Lim, "Abstractive Summarization of COVID-19 Medical Research Articles using BERT and GPT-2," *Journal of Medical Systems*, vol. 44, no. 12, 2020.

[34] N. Sangavi, M. Umamaheswari, and V. Subasri, "NLP Based Text Summarization Using BART Model," *International Journal of Scientific Research in Engineering and Management*, 2023.

[35] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," 2022. doi: 10.1007/s00521-022-07366-3.

[36] J. Krantz and J. Kalita, "Abstractive Summarization Using Attentive Neural Techniques," 2018. Available: https://arxiv.org/abs/1810.08838.

[37] T. J. Ham, Y. Lee, S. H. Seo, S.-U. Kim, H. Choi, S. Jung, and J. W. Lee, "ELSA: Hardware-Software Co-design for Efficient, Lightweight Self-Attention Mechanism in Neural Networks," 2021. doi: 10.1109/ISCA52012.2021.00060.

[38] S. Tang, J. Zhang, S. Zhu, and P. Tan, "QuadTree Attention for Vision Transformers," *ArXiv*, abs/2201.02767, 2018.

[39] W. Yang, Z. Tang, and X. Tang, "A Hierarchical Neural Abstractive Summarization with Self-Attention Mechanism," *166(Amcce)*, pp. 514–518, 2018. doi: 10.2991/amcce-18.2018.89.

[40] X. Duan, H. Yu, M. Yin, M. Zhang, W. Luo, and Y. Zhang, "Contrastive Attention Mechanism for Abstractive Sentence Summarization," *ArXiv*, abs/1910.13114, 2019.

[41] Z. Wang, "An Automatic Abstractive Text Summarization Model based on Hybrid Attention Mechanism," 2021. doi: 10.1088/1742-6596/1848/1/012057.

[42] H. Li, S. Xu, P. Yuan, Y. Wang, Y. Wu, X. He, and B. Zhou, "Learn to Copy from the Copying History: Correlational Copy Network for Abstractive Summarization," 2021. doi: 10.18653/v1/2021.emnlp-main.336.

[43] Q. Zhou, N. Yang, F. Wei, and M. Zhou, "Sequential Copying Networks," 2018. doi: 10.1609/aaai.v32i1.11915.

[44] S. Ren and Z. Zhang, "Pointer-Generator Abstractive Text Summarization Model with Part of Speech Features," 2019. doi: 10.1109/ICSESS47205.2019.9040715.

[45] Z. Liu, A. Ng, S. S. G. Lee, A. Aw, and N. F. Chen, "Topic-Aware Pointer-Generator Networks for Summarizing Spoken Conversations," 2019. doi: 10.1109/ASRU46091.2019.9003764.

[46] R. Nallapati, B. Zhou, C. N. Santos, Ç. Gülçehre, and B. Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in *Conference on Computational Natural Language Learning*, 2016.

[47] L. Chen, Y. Zhang, R. Zhang, C. Tao, Z. Gan, H. Zhang, B. Li, D. Shen, and C. Chen, "Improving Sequence-to-Sequence Learning via Optimal Transport," 2019.

[48] P. Fecht, S. Blank, and H.-P. Zorn, "Sequential Transfer Learning in NLP for German Text Summarization," in *3rd International Seminar on Education Innovation and Economic Management (SEIEM 2018)*, 2019. doi: 10.2991/SEIEM-18.2019.131.

[49] E. Zolotareva, T. M. Tashu, and T. Horváth, "Abstractive Text Summarization using Transfer Learning," *IEEE Access*, 2020.

[50] S. Buciumas, "Reinforcement Learning Models for Abstractive Text Summarization," *ACM Digital Library*, 2019.

[51] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies," in *North American Chapter of the Association for Computational Linguistics*, 2018.

[52] A.M. Rush, S. Chopra, and J. Weston, "A Neural Attention Model for Abstractive Sentence Summarization," in *Conference on Empirical Methods in Natural Language Processing*, 2015.

[53] Y. Zhang, Y. Liu, Z. Yang, Y. Fang, Y. Chen, D. R. Radev, C. Zhu, M. Zeng, and R. Zhang, "MACSum: Controllable Summarization with Mixed Attributes," *Transactions of the Association for Computational Linguistics*, 2023.

[54] S. Narayan, S. B. Cohen, and M. Lapata, "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization of Source Code," *arXiv preprint arXiv:1808.08745*, 2018.

[55] C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Annual Meeting of the Association for Computational Linguistics*, 2004.

[56] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Annual Meeting of the Association for Computational Linguistics*, 2002.

[57] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *IEEvaluation@ACL*, 2005.

[58] N. Lin, J. Li, and S. Jiang, "A simple but effective method for Indonesian automatic text summarisation," *Connection Science*, vol. 34, no. 1, pp. 29–43, 2022. doi: 10.1080/09540091.2021.1937942.

[59] I. S. Blekanov, N. Tarasov, and S. S. Bodrunova, "Transformer-Based Abstractive Summarization for Reddit and Twitter: Single Posts vs. Comment Pools in Three Languages," *Future Internet*, vol. 14, no. 3, p. 69, 2022. doi: 10.3390/fi14030069.

[60] Y. Zhao, M. Khalman, R. Joshi, S. Narayan, M. Saleh, and P.J. Liu, "Calibrating Sequence likelihood Improves Conditional Language Generation," *ArXiv*, abs/2210.00045, 2022.

[61] B. He, J. Wang, J. Qiu, T. Bui, A. Shrivastava, and Z. Wang, "Align and Attend: Multimodal Summarization with Dual Contrastive Losses," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14867-14878, 2023.

[62] F. Wang, K. Song, H. Zhang, L. Jin, S. Cho, W. Yao, X. Wang, and M. Chen, "Salience Allocation as Guidance for Abstractive Summarization," *ArXiv*, abs/2210.12330, 2022.

[63] M. Ravaut, S. Joty, and N. F. Chen, "SummaReranker: A Multi-Task Mixture-of-Experts Re-ranking Framework for Abstractive Summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 2022. doi: 10.18653/v1/2022.acl-long.309.

[64] A. Afzal, J. Vladika, D. Braun, and F. Matthes, "Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them," in *International Conference on Agents and Artificial Intelligence*, 2023. doi: 10.5220/0011744500003393.

[65] M. Bonnaerens and J. Dambre, "Learned Thresholds Token Merging and Pruning for Vision Transformers," *ArXiv*, abs/2307.10780, 2023.

[66] Y. Wu, S. Kan, M. Zeng, and M. Li, "Singularformer: Learning to Decompose Self-Attention to Linearize the Complexity of Transformer," in *International Joint Conference on Artificial Intelligence*, 2023.

[67] F. Nie, J.-g. Yao, J. Wang, R. Pan, and C.-Y. Lin, "A Simple Recipe towards Reducing Hallucination in Neural Surface Realisation," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[68] D. Lei, Y. Li, M. Hu, M. Wang, V. Yun, E. Ching, and E. Kamal, "Chain of Natural Language Inference for Reducing Large Language Model Ungrounded Hallucinations," *ArXiv*, 2023. Available: https://arxiv.org/abs/2104.14839.

[69] V. Srivastava, S. Bhat, and N. Pedanekar, "Hiding in Plain Sight: Insights into Abstractive Text Summarization," *ArXiv*, abs/2104.14839, 2023.

[70] P.A. Utama, J. Bambrick, N.S. Moosavi, and I. Gurevych, "Falsesum: Generating Document-level NLI Examples for Recognizing Factual Inconsistency in Summarization," *ArXiv*, abs/2205.06009, 2022.

[71] J. Sul and Y.S. Choi, "Balancing Lexical and Semantic Quality in Abstractive Summarization," *ArXiv*, abs/2305.09898, 2023.

[72] G. Adams, B.H. Nguyen, J. Smith, Y. Xia, S. Xie, A. Ostropolets, B. Deb, Y. Chen, T. Naumann, and N. Elhadad, "What are the Desired Characteristics of Calibration Sets? Identifying Correlates on Long Form Scientific Summarization," in *Annual Meeting of the Association for Computational Linguistics*, 2023.

[73] Y.-C. Huang, X. Feng, X. Feng, and B. Qin, "The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey," *ArXiv*, 2021. Available: https://arxiv.org/abs/2104.14839.

[74] Z. Luo, Q. Xie, and S. Ananiadou, "ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization," *ArXiv*, abs/2303.15621, 2023.

[75] Y. Liang, F. Meng, J. Wang, J. Xu, Y. Chen, and J. Zhou, "D2TV: Dual Knowledge Distillation and Target-oriented Vision Modeling for Many-to-Many Multimodal Summarization," in *Conference on Empirical Methods in Natural Language Processing*, 2023.

[76] N. Shafiq, I. Hamid, M. Asif, Q. Nawaz, H. Aljuaid, and H. Ali, "Abstractive text summarization of low-resourced languages using deep learning," *PeerJ Computer Science*, 2023. doi: 10.7717/peerj-cs.1176.

[77] L. Mascarell, R. Chalumattu, and J. Heitmann, "Entropy-based Sampling for Abstractive Multi-document Summarization in Low-resource Settings," in *International Conference on Natural Language Generation*, 2023.

[78] T. Hasan, A. Bhattacharjee, M. S. Islam, K. S. Mubasshir, Y.-F. Li, Y.-B. Kang, M. Rahman, and R. Shahriyar, "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages," in *Findings*, 2021.

[79] A. Dash, A. Shandilya, A. Biswas, A. Chakraborty, K. Ghosh, and S. Ghosh, "Beyond ROUGE Scores in Algorithmic Summarization: Creating Fairness-Preserving Textual Summaries," *ArXiv*, abs/1810.09147, 2018.

[80] Y. Gao, W. Zhao, and S. Eger, "SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization," *ArXiv*, abs/2005.03724, 2020.

[81] I. Saito, K. Nishida, K. Nishida, A. Otsuka, H. Asano, J. Tomita, H. Shindo, and Y. Matsumoto, "Length-controllable abstractive summarization by guiding with summary prototype," *ArXiv*, abs/2005.12345, 2020.

[82] D. Fein and R. Cuevas, "ExtraPhraseRank: A Length-Controlled Data Augmentation Strategy for Unsupervised Abstractive Summarization," 2022. Available: https://arxiv.org/abs/2012.03656.

[83] M. Y. Abdelwahab, Y. Al Moaiad, and Z. Abu Bakar, "Arabic Text Summarization Using Pre-Processing Methodologies and Techniques," Asia-Pacific Journal of Information Technology and Multimedia, vol. 12, no. 1, pp. 70-110, 2023. Available: http://dx.doi.org/10.17576/apjitm-2023-1201-05.

[84] E. Heidary, H. Parvïn, S. Nejatian, K. Bagherifard, V. Rezaie, Z. Mansor, and K.-H. Pho, "Automatic Text Summarization Using Genetic Algorithm and Repetitive Patterns," Computers, Materials & Continua, Tech Science Press, 2021. Available: https://doi.org/10.32604/cmc.2021.013836.

[85] S. B. bin Rodzman, N. K. Ismail, N. A. Rahman, S. A. Aljunid, H. A. Rahman, Z. M. Nor, K. M. Khalif, and A. Y. M. Noor, "Experiment with Text Summarization as a Positive Hierarchical Fuzzy Logic Ranking Indicator for Domain Specific Retrieval of Malay Translated Hadith," 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Malaysia, 2019, pp. 299-304. Available: https://doi.org/10.1109/ISCAIE.2019.8743988.

[86] H. N. Fejer and N. Omar, "Automatic Arabic text summarization using clustering and keyphrase extraction," Proceedings of the 6th International Conference on Information Technology and Multimedia, Putrajaya, Malaysia, 2014, pp. 293-298. Available: https://doi.org/10.1109/ICIMU.2014.7066647.

[87] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "An Optimized LSTM-Based Augmented Language Model (FLSTM-ALM) Using Fox Algorithm for Automatic Essay Scoring Prediction," in *IEEE Access*, vol. 12, pp. 48713-48724, 2024. Available: https://doi.org/10.1109/ACCESS.2024.3381619.