




- ORIGINAL ARTICLE -

# Automatic Text Summarization: A Review of Approaches, Challenges, and Future Directions

## Resumen automático de textos: una revisión de enfoques, desafíos y direcciones futuras

Sara Zayed <sup>1</sup> , Mostafa Ezzat <sup>1</sup>  and Hesham A. Hefny <sup>1</sup> 

<sup>1</sup> *Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo University, Cairo, Egypt*

{sarazayed@cu.edu.eg; mostafa.ezzat@cu.edu.eg; hehefny@cu.edu.eg}

### Abstract

Automatic text summarization (ATS) is a vital area in natural language processing focused on condensing lengthy documents into concise, meaningful summaries. Manual summarization remains time-consuming and costly, motivating extensive research into extractive, abstractive, and hybrid methods. This review provides a comprehensive survey of ATS, covering traditional techniques alongside recent breakthroughs fueled by machine learning, deep learning, and transformer-based architectures, including large language models (LLMs) such as GPT-4, Claude, Falcon, and MPT. We introduce a novel multi-dimensional taxonomy that integrates classical methods, emerging paradigms, domain-specific and multimodal approaches, and instruction-tuned LLMs. Our survey analyzes the strengths and limitations of each approach and discusses evaluation methodologies, incorporating both established metrics like ROUGE and BLEU and newer ones. We further address key challenges including factual consistency, prompt sensitivity, explainability, ethical considerations, and computational efficiency. By bridging traditional summarization techniques with cutting-edge LLM-based models, this paper highlights current limitations, emerging opportunities, and future directions aimed at developing robust, reliable, and scalable summarization systems for diverse real-world applications.

**Keywords:** Automatic Text Summarization, Text Summarization, Taxonomy, Transformers, LLMs.

### Resumen

El resumen automático de texto (ATS) es un área vital del procesamiento del lenguaje natural, centrada en condensar documentos extensos en resúmenes concisos y significativos. El resumen manual sigue siendo lento y costoso, lo que motiva

una amplia investigación sobre métodos extractivos, abstractos e híbridos. Esta revisión ofrece un estudio exhaustivo de los ATS, que abarca las técnicas tradicionales junto con los avances recientes impulsados por el aprendizaje automático, el aprendizaje profundo y las arquitecturas basadas en transformadores, incluyendo grandes modelos de lenguaje (LLM) como GPT-4, Claude, Falcon y MPT. Presentamos una novedosa taxonomía multidimensional que integra métodos clásicos, paradigmas emergentes, enfoques multimodales y específicos de cada dominio, y LLM optimizados para instrucciones. Nuestro estudio analiza las fortalezas y limitaciones de cada enfoque y analiza las metodologías de evaluación, incorporando tanto métricas consolidadas como ROUGE y BLEU como otras más recientes. Además, abordamos desafíos clave como la coherencia fáctica, la sensibilidad inmediata, la explicabilidad, las consideraciones éticas y la eficiencia computacional. Al combinar las técnicas de resumen tradicionales con modelos de vanguardia basados en LLM, este documento destaca las limitaciones actuales, las oportunidades emergentes y las direcciones futuras destinadas a desarrollar sistemas de resumen sólidos, confiables y escalables para diversas aplicaciones del mundo real.

**Palabras claves:** Resumen automático de texto, resumen de texto, taxonomía, transformadores, LLMs.

### 1. Introduction

Recently, Automatic text summarization has witnessed an exponential growth as an essential tool for efficiently condensing large amounts of text into shorter, coherent summaries. Manual summarization is labor-intensive, time-consuming, and expensive, which highlights the growing importance of automated methods [1] [2]. With the massive growth of textual data across various sources and domains, it has become increasingly difficult for humans to manually summarize this massive information [3]. Automatic text summarization is a key solution to

this problem [3], providing the ability to automate the process of condensing large texts into concise, informative summaries.

Automatic text summarization works through two main approaches: extractive summarization, and abstract summarization [4]. Recent advances in machine learning, particularly with transformer-based models, have greatly improved the quality of summaries. Despite these advances, challenges remain, such as ensuring coherence, avoiding redundancy, and preventing information loss.

In extracted text summarization, the resulting summary consists of precise phrases and sentences from the original text. The algorithm selects key parts of the content and combines them to form a coherent summary. In contrast, abstract text summarization generates a summary that may or may not use the same phrases as the original text. Instead, the algorithm interprets the meaning and rephrases the content, resulting in a more flexible and human-like summary [1].

The formal definition of automatic text summarization is given in the book [5], where it is described as “the process of extracting the most important information from a source (or different sources) to produce a condensed version for a particular user (or users) and task (or tasks)”.

Due to the great interest in the field, several papers have been published that explore various approaches and methods to develop automated solutions for text summarization. As part of our study, we focused on studies published over recent years related to the automation of text summarization.

While several previous surveys [57–60] have provided valuable insights into automatic text summarization, our work offers several important distinctions. Earlier reviews mainly focus on traditional extractive and abstractive methods, application scenarios, and broad research trends. However, they either do not deeply address the rapid evolution brought by transformer-based models and large language models or only mention them briefly.

Our survey explicitly incorporates a detailed exploration of transformer architectures and LLM-based summarization techniques, covering both extractive and abstractive methods. Additionally, unlike prior works that primarily analyze studies up to 2019 or 2020, we provide a more updated perspective, highlighting recent breakthroughs and the shift towards LLM-driven summarization. Unlike these surveys, this review provides an updated and comprehensive view of the field in the LLM era. It offers a new taxonomy encompassing transformer variants, multimodal approaches, and domain-specific adaptations, alongside practical mappings between model architectures, datasets, and application domains. Additionally, this survey uniquely addresses emerging challenges such as

factual consistency in LLM outputs, scalable long-input handling, and domain adaptation, serving as a critical resource for guiding future summarization research. Additionally, we discussed emerging research in low-resource languages like Arabic, Hindi, etc., and we discussed emerging challenges, such as factual consistency and domain-specific summarization, which are becoming increasingly important in the LLM era. By offering this expanded and updated view, our survey serves as a comprehensive and timely resource for researchers and practitioners aiming to navigate the evolving landscape of ATS.

Therefore, our goal is to help researchers in the natural language processing (NLP) field understand the extent of the problem, evaluate the effectiveness of existing models, and develop customized solutions to advance the field.

The main contributions of this survey are as follows: (1) it introduces an expanded taxonomy that goes beyond the traditional extractive, abstractive, and hybrid categories to include transformer-based models, multimodal and cross-lingual approaches, and LLM-based summarization systems; (2) it provides an in-depth analysis of instruction-tuned LLMs like GPT-4 and Claude, highlighting new issues such as factual consistency, hallucinations, and prompt engineering, which have not been thoroughly discussed in earlier reviews; (3) it proposes a novel task–model–dataset mapping framework that links summarization types, model architectures, domains, and benchmark datasets to guide practical deployment; (4) it presents an updated review of hybrid summarization techniques, focusing on recent extract-then-abstract pipelines enhanced with methods like topic modeling, fuzzy logic, and optimization for better performance on long or domain-specific inputs; and (5) it outlines key research gaps and future directions, including the need for better multilingual datasets, more robust evaluation metrics, and progress toward controllable, fact-aware, and explainable summarization systems in the post-LLM era.

## 2. Methodology

The main objective of this study is to investigate the state of the art of the latest techniques to automatically summarize texts. This survey covers the following research questions:

- 1) What are the key approaches to ATS, including traditional methods, transformer-based models, and large language models, and how do they differ in terms of methodology, capabilities, and effectiveness?
- 2) What are the essential building blocks, architectural paradigms, and emerging techniques

such as instruction tuning, prompt engineering, and hybrid extract-then-abstract pipelines used in developing effective ATS systems?

3) What datasets are commonly utilized for training and evaluating ATS models, including both benchmark and domain-specific datasets, and what are their characteristics in terms of content, language, and task type?

4) What evaluation metrics are most effective for assessing the performance of traditional and LLM-based ATS methods, and how do they impact the interpretation of results, especially regarding factual consistency and summary quality?

5) What are the current challenges and limitations faced in ATS, including dataset scarcity in low-resource languages, hallucinations in LLMs, and the inadequacy of existing evaluation metrics, and how can these be addressed in future research?

6) How have recent advancements in deep learning, transformer architectures, and LLMs influenced the development, generalization, and real-world applicability of ATS techniques?

7) What are the future research directions in ATS, including the need for explainability, controllability, domain adaptation, and cross-lingual summarization, and how can these enhance the field's contribution to intelligent information processing?

This review follows a systematic approach to gather and analyze relevant literature on Automatic Text Summarization. The methodology encompasses the following steps:

## 2.1. Search strategy

A comprehensive literature search was conducted across various academic databases, including IEEE, ACL Anthology, MDPI, IJITEE, EDP Sciences, COAS, ScienceDirect, ArXiv, Nature portfolio and Springer. We also focused on recent works, especially from top NLP venues, which have pushed the boundaries of abstractive summarization. The search included keywords such as "automatic text summarization," "extractive summarization," "abstractive summarization," "evaluation metrics," and "datasets." The search was limited to publications from the last decade to ensure the inclusion of recent advancements in the field.

### 2.1.1. Inclusion and Exclusion Criteria

The inclusion criteria focused on peer-reviewed articles, conference papers, and relevant book chapters that address various aspects of Automatic Text Summarization, including methodologies, techniques, applications, and challenges. We prioritized journal papers with a high impact factor, followed by high-quality conference proceedings. Papers not published in English, lacking substantial content related to ATS, or with insufficient

methodological rigor were excluded. Additionally, papers that did not meet peer-review standards or lacked detailed experimental validation were also excluded from the review.

### 2.1.2. Data Extraction and Organization

Relevant information was extracted from the selected papers, including the approach used (extractive or abstractive), the algorithms employed, datasets utilized, evaluation metrics applied, and key findings. This information was organized in a structured format, allowing for a clear comparison of different studies.

### 2.1.3. Analysis and Synthesis

The data collected were analyzed to identify trends, common challenges, and gaps in the current research landscape. The synthesis of findings aimed to highlight the strengths and limitations of existing approaches, as well as to suggest future research directions for improving ATS techniques.

### 2.1.4. Critical Evaluation

A critical evaluation of the methodologies employed in the reviewed studies was conducted to assess their effectiveness and applicability. This evaluation included considerations of the algorithms' performance, the quality of datasets, and the robustness of evaluation methods.

Finally, after a thorough analysis, we focused on recent papers to ensure our review reflects the latest advancements in the field and continues to push the boundaries of research. However, we also included earlier works to ensure that the contributions and efforts of previous researchers are not overlooked. This approach resulted in the inclusion of several recent studies in our survey from various databases published over recent years, as shown in Fig. 1 and Fig. 2 respectively.

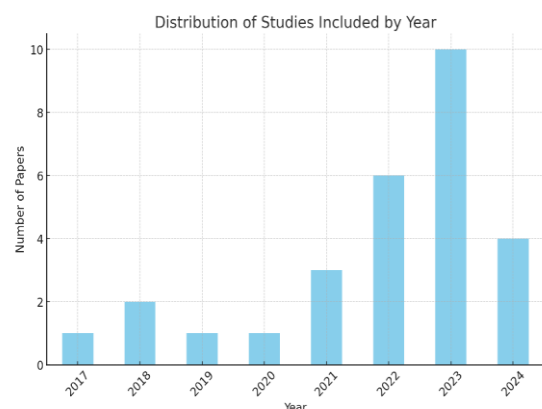


Fig. 1 Distribution of selected studies per year.

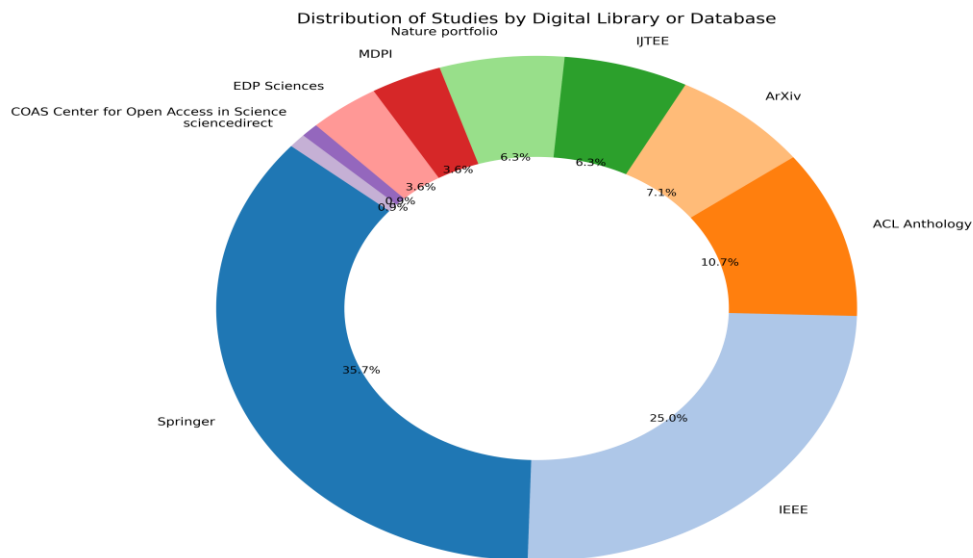


Fig. 2 Distribution of selected studies per digital library.

This survey was conducted to provide a background on automatic text summarization by answering the questions mentioned above. The rest of this paper is organized as follows: The subsequent sections will provide a background of the distinct types of ATS, key building blocks, relevant datasets, and evaluation metrics. Additionally, current challenges faced in the field will be discussed, along with recent advancements and future research directions. This structure highlights the evolving landscape of ATS and its potential to address the growing demand for efficient information processing in the digital age.

### 3. Taxonomy: Types of Automatic Text Summarization

Automatic Text Summarization can be classified into three main types: Extractive Summarization, Abstractive Summarization, and Hybrid Summarization.

#### 3.1 Traditional Classification of Summarization Approaches

##### 3.1.1 Extractive Summarization

In extractive text summarization, the model identifies and selects important sentences from the source document to form the summary [6]. It treats summarization as a classification problem, where important sentences are directly selected from the source text to construct a summary [7]. Each sentence is assigned a weight, and the highest-ranked sentences are extracted and combined to

generate the summary. While these summaries are often less coherent, extractive methods are widely used due to their lower time complexity and ease of generation compared to abstractive summarization. A good extractive summary should have topic diversity and low redundancy, though balancing these aspects is challenging [6].

##### 3.1.2 Abstractive Summarization

In contrast to extractive methods, abstractive summarization method utilizes neural network-based approaches, such as the Sequence-2-Sequence (Seq2Seq) architecture [7]. The model processes the text, rephrases information, and produces a condensed version in its own words. This approach mimics how humans summarize and can lead to more coherent and readable results. However, abstractive summarization requires advanced language understanding and generation capabilities, making it more complex and prone to errors or information distortion, especially with shorter summaries. It produces summaries with less redundancy but may struggle to maintain fluency and grammatical accuracy [7].

##### 3.1.3 Hybrid Summarization

Hybrid summarization is a fusion of extractive and abstractive techniques [13]. It combines elements of both extractive and abstractive approaches. In this method, key sentences or passages may first be extracted from the text, and then, these selected elements undergo further refinement or rephrasing to improve coherence and readability. By leveraging

the strengths of both methods, hybrid summarization aims to maintain factual accuracy (from the extractive approach) while enhancing fluency and naturalness (from the abstractive approach). This type is increasingly gaining traction, as it balances simplicity with the generation of more natural summaries, especially in tasks that require both accuracy and readability.

Finally, we performed a taxonomy analysis of the various techniques used in Automatic Text Summarization, focusing on the above methods. Each method incorporates distinct techniques to achieve effective summarization as follows:

- a. Extractive methods include a wide range of approaches such as Statistical models, Topic-Based methods, Clustering, Graph-Based techniques, Semantic analysis, Machine Learning, Neural Networks, Fuzzy Logic, Deep Learning, and Optimization techniques. These approaches focus on selecting important segments from the original text to create summaries.
- b. Abstractive methods rely on more advanced techniques like Graph-Based structures, Semantic approaches, Deep Learning models, Domain-Specific methods, and Tree-based representations. These techniques generate summaries by rephrasing and reinterpreting the content, mimicking human summarization processes.
- c. Hybrid methods combine aspects of both Extractive and Abstractive approaches, such as transitioning from Extractive methods to shallow Abstractive summarization, to benefit from both factual accuracy and linguistic coherence.

From our analysis of the included papers, we identified a diverse distribution of techniques used across various summarization approaches. Extractive methods are the most prevalent, comprising half of the studies, while abstract and hybrid methods make up the rest. Among the techniques, semantic and topic-based methods are the most frequently used, each appearing in multiple studies, followed by domain-specific approaches, machine learning, and fuzzy logic, which also have notable representation. Less common techniques include graph-based, neural networks, and statistical methods, each applied in one or two studies.

This distribution highlights the breadth of methods explored in recent research for text summarization, showcasing advancements in both traditional and modern approaches. Fig.3 and Fig.4 visualize the distribution of ATS types and techniques (topic/trend).

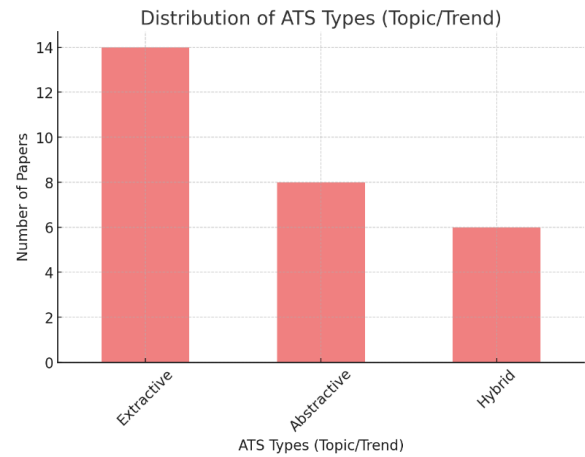


Fig.3 Distribution of Traditional Classification of Summarization Approaches.

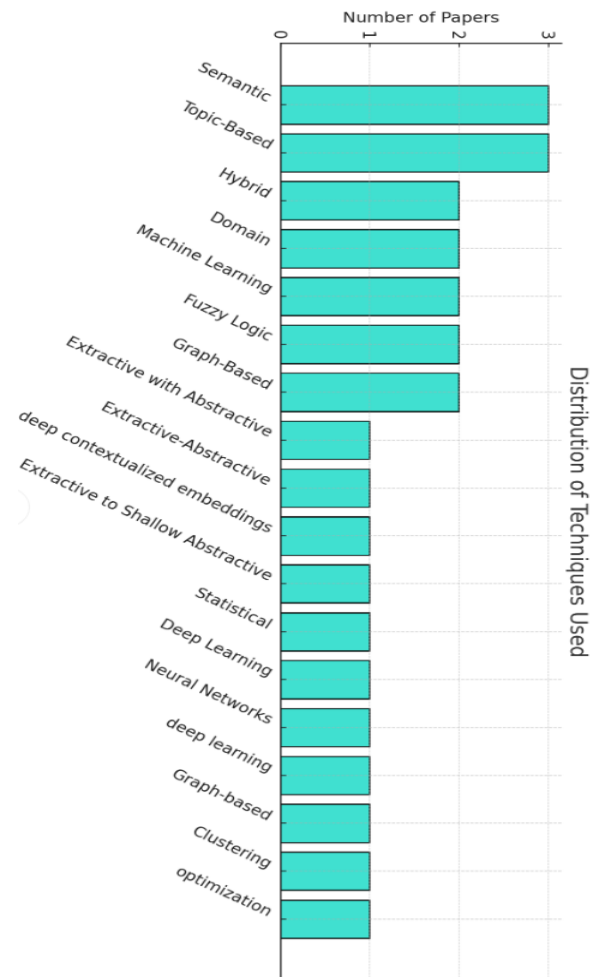


Fig.4 Distribution of ATS techniques used over literature.

### 3.2 Expansion to Multi-Dimensional Summarization Paradigms

Recent advances demand a broader taxonomy that accounts for new dimensions beyond the classical three. We identify four emerging paradigms:

### a. Transformer-Based Summarization

Transformers architectures have significantly advanced the field of automatic text summarization by enabling models to capture complex dependencies, contextual semantics, and long-range relationships across input sequences. Unlike traditional RNN- or CNN-based models, transformers utilize self-attention mechanisms that allow parallel processing of tokens, enhancing both efficiency and contextual understanding [73]. Several notable transformer-based models have become the backbone of modern summarization systems:

- BART (Bidirectional and Auto-Regressive Transformers) combines a denoising autoencoder with a Seq2Seq architecture, proving highly effective for both abstractive and extractive summarization tasks [74].
- PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) introduces a novel pre-training objective tailored specifically for summarization, outperforming earlier models on multiple benchmarks [75].
- Longformer and BigBird are designed to handle longer documents by replacing the standard quadratic attention mechanism with sparse attention, enabling efficient processing of inputs exceeding 4,000 tokens [76,77].
- LongT5, an extension of the T5 model, integrates sparse attention with improved encoder-decoder design, facilitating summarization of extensive legal or scientific texts with higher efficiency [78].
- PEGASUS-X further extends PEGASUS by incorporating multi-query attention and optimized scaling strategies, allowing it to manage long-form summarization tasks in domains such as biomedical and legal documentation [79].

These transformer-based approaches are increasingly applied in specialized domains like legal, medical, and financial summarization, where input documents tend to be significantly longer and more complex than standard datasets like CNN/DailyMail. Their architecture and pre-training tasks make them especially suited for capturing nuanced content, maintaining coherence, and reducing redundancy in summaries.

### b. Multimodal and Multilingual Summarization

Multimodal summarization integrates text with other modalities such as images (e.g., news articles), audio transcripts (e.g., meetings), or videos (e.g., YouTube, lectures). Models like Flamingo and Kosmos [80, 81] handle multi-input formats, expanding summarization beyond pure text. Meanwhile, multilingual and cross-lingual summarization addresses language diversity, enabling summarization of content in

underrepresented languages or translating summaries across languages using models like mBART and mT5.

### c. Domain-Specific Summarization

Specialized summarization systems are designed for domains such as biomedical (PubMed), legal (BillSum), financial (BioBART [82], FinLlama3 [83]), and conversational (SAMSum). These models often require domain-adaptive pretraining or fine-tuning and utilize tailored vocabularies, knowledge bases, or constraints to ensure summary relevance and correctness.

### d. Large Language Model (LLM)-Based Summarization

Instruction-tuned large language models (LLMs) such as GPT-4, Claude, Mistral, and LLaMA are reshaping automatic summarization by enabling zero-shot, few-shot, and prompt-driven summarization approaches. These models demonstrate strong generalization across domains and tasks, significantly reducing the need for task-specific fine-tuning. However, they also introduce challenges related to controllability, factual consistency, and evaluation reliability. Unlike conventional models, LLMs can generate summaries with minimal training data, making them attractive for rapid deployment in real-world scenarios. Studies have shown that models like GPT-3.5 and GPT-4 outperform traditional transformer models on datasets like CNN/DailyMail and SAMSum, especially in coherence and informativeness, but may still suffer from hallucinations and poor factual grounding in domain-specific contexts [84].

## 3.3 Technique-Based Classification Across Dimensions

Our analysis also categorizes summarization techniques based on the algorithms used across extractive, abstractive, and hybrid systems:

- **Extractive:** Statistical models, topic modeling, clustering, graph-based algorithms (e.g., LexRank [85]), semantic similarity, optimization techniques (e.g., genetic algorithms), neural scoring, and fuzzy logic.
- **Abstractive:** Transformer-based Seq2Seq models, reinforcement learning, copy mechanisms, semantic parsing, tree-based representations, and domain-specific fine-tuning. The introduction of models like BART and PEGASUS has significantly advanced neural abstractive summarization [75].
- **Hybrid:** Pipeline models (e.g., extract-then-generate), margin ranking, knowledge-enhanced

summarization, topic-guided abstraction, and fuzzy-rule driven re-ranking [86].

Our taxonomy not only revisits the classical ATS types but also extends into a multi-dimensional space that accounts for model architecture, input modality, domain specialization, and task complexity. This provides a structured lens through which future research can be contextualized. Figure 5 presents an integrated taxonomy that maps summarization types to techniques, domains, and model categories.

#### 4. Preprocessing Techniques and Feature Extraction

Preprocessing is a main and important task in NLP [56]. Text summarization requires robust preprocessing and feature extraction to ensure that only the most important information is captured from the source text. Preprocessing prepares the data by removing irrelevant information, while feature extraction identifies informational features that are important for summarization models.

##### 4.1. Preprocessing Techniques

- 1) **Tokenization:** Text is split into tokens which are normally words. Tokenization is foundational for further text processing and is often done using rule-based or language-specific tokenizers, particularly to address differences in punctuation and syntax across languages [26]. Tokenization has been extensively used in summarization tasks, as outlined by [6] in their research on deep neural network-based summarization models to reduce the complexity of words in documents.
- 2) **Stopwords Removal:** Commonly used words (e.g., “and” “the”) are removed because they do not carry meaningful information for summarization purposes. Removing stopwords reduces the dataset’s dimensionality and focuses on the most informative terms [1] [6].
- 3) **Stemming and Lemmatization:** These techniques are both used to reduce words to their root forms, but they differ in their approach. Stemming involves chopping off prefixes or suffixes from words to get to a root form (e.g., “running” to “run”), but the result may not always be a valid word. On the other hand, lemmatization reduces words to their base or dictionary form (e.g., “running” to “run”) by considering the word’s meaning and context, ensuring that the resulting word is a valid word in the language. Both techniques help to avoid redundancy and group similar concepts together, focusing on the core meaning [6], which is crucial in summarizing content

accurately. However, stemming might not always significantly impact performance [13].

- 4) **Lowercasing and Normalization:** Converting text to lowercase and handling other linguistic variations like spelling differences or contractions can standardize the text, which is essential when dealing with large datasets. This is often combined with additional preprocessing for handling punctuation and special characters. The authors in [13] converted all letters to lower-case for the model not to differentiate between words in the beginning and in the middle of sentences, such as “apples” and “Apples”.

##### 4.2. Feature Extraction Methods

Feature extraction identifies the key words and phrases essential for understanding the core meaning of the text being analyzed [25]. These features can be broadly classified into linguistic, statistical, and neural-based features.

###### 4.2.1. Linguistic Features:

Linguistic features include parts of speech (POS) tags, named entities, and syntactic structures. POS tagging helps identify nouns, verbs, adjectives, and other parts of speech to capture the main concepts, while named entity recognition (NER) helps highlight important entities like names, dates, and locations. Linguistic features are often applied in extractive summarization to retain meaningful sentences [27].

###### 4.2.2. Statistical Features:

Statistical methods identify terms based on their frequency and position. Common statistical features include:

- 1) **Term Frequency-Inverse Document Frequency (TF-IDF):** This method calculates the importance of terms in a document relative to the corpus, favoring terms that occur frequently in a document but not in others [28].
- 2) **Sentence Position:** Sentences at the beginning or end of paragraphs often contain summaries or conclusions. Assigning weights to these sentences can help improve summarization accuracy [29].
- 3) **Sentence Length and Similarity:** Filtering out very short sentences (or overly long ones) and using similarity measures (e.g., cosine similarity) between sentences help identify sentences that carry essential information while avoiding redundancy [30].

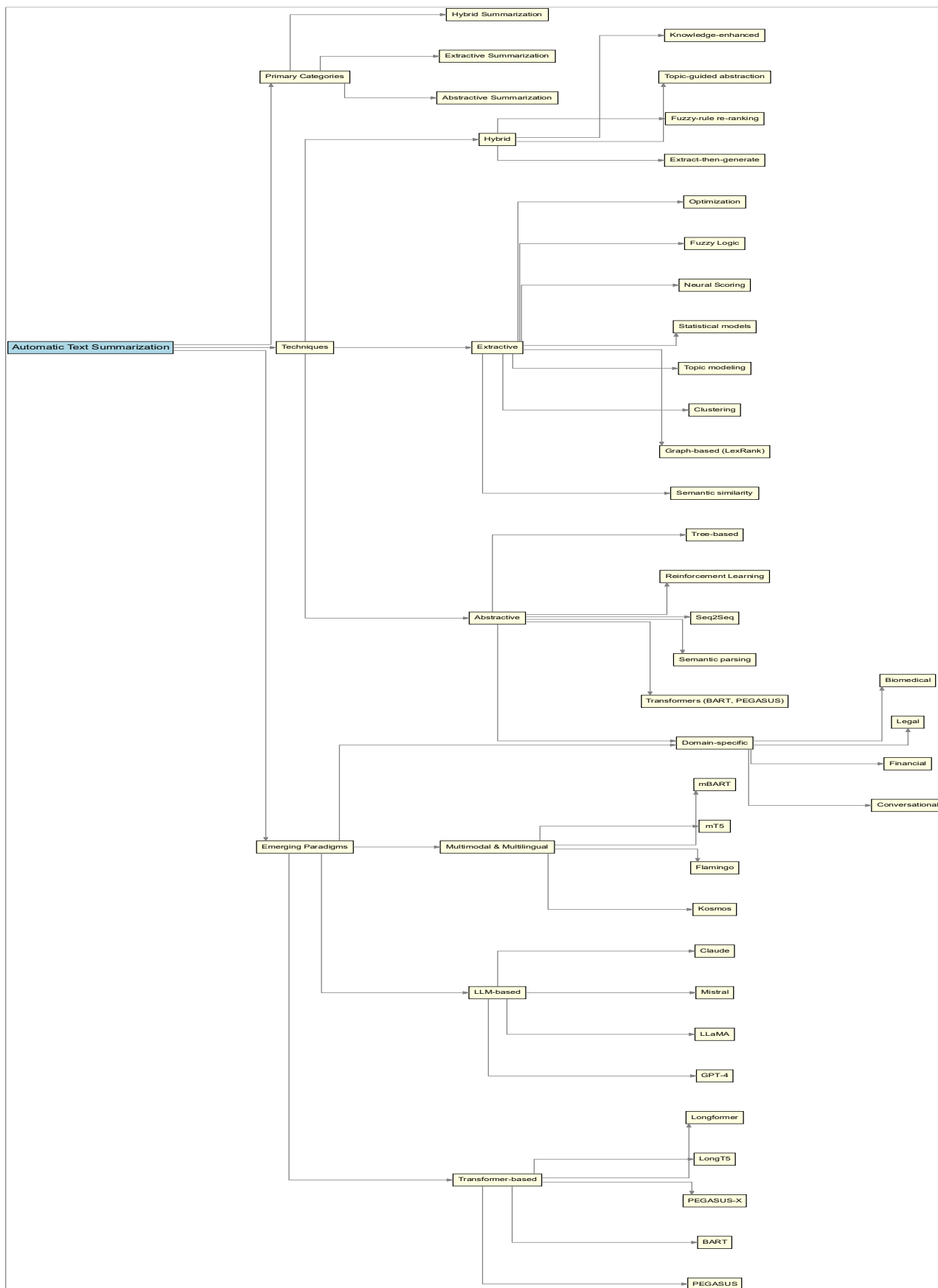


Fig.5 A taxonomy analysis of summarization types, techniques, domains, and model categories.

#### 4.2.3. Neural-Based Features:

With advances in deep learning, neural-based approaches have become prominent for feature extraction in text summarization. Pretrained transformers, like BERT and its variants, are used to extract contextual embeddings that capture the semantic richness of sentences and words. These embeddings are used to generate summaries by identifying sentence importance based on context rather than term frequency or statistical measures. Transformers have shown remarkable improvements in abstractive summarization, where the summary is generated using natural language generation [31].

## 5. Large Language Models in Summarization

### 5.1. Introduction to LLM-Based Summarization

The emergence of Large Language Models, such as GPT-4 [87], Claude [88], Falcon [89], and MPT [90], has fundamentally transformed automatic text summarization. These models, based on transformer architectures with billions of parameters, excel at capturing deep contextual and semantic information to generate coherent and fluent natural language text. Beyond classical summarization models which often require supervised fine-tuning on domain-specific corpora, LLMs demonstrate remarkable abilities for zero-shot and few-shot summarization through instruction tuning and prompt engineering, enabling adaptation to diverse summarization tasks with minimal task-specific training [91].

### 5.2. Instruction-Tuned LLMs for Summarization

Instruction tuning has emerged as a pivotal technique for enhancing the summarization capabilities of large language models (LLMs). By fine-tuning models on datasets comprising instruction-output pairs, LLMs can better align with human intent and produce more coherent and contextually relevant summaries. Notably, instruction-tuned models such as ChatGPT (based on GPT-4), Falcon-7B, MPT-7B, and Claude have demonstrated significant improvements in summarization tasks. For example, ChatGPT has shown strong performance in extractive summarization when used in an extract-then-generate pipeline, improving factual consistency and fluency [91]. Similarly, Falcon-7B and MPT-7B have been evaluated on benchmark datasets like CNN/DailyMail and XSum, exhibiting competitive performance in zero-shot and few-shot settings [68]. These advancements underscore the efficacy of

instruction tuning in enabling LLMs to perform summarization tasks without extensive task-specific fine-tuning, thereby broadening their applicability across diverse domains and summarization styles.

## 6. Datasets for ATS

In the field of automatic text summarization, datasets play a crucial role in developing and evaluating summarization models. High-quality datasets ensure that models are trained effectively to generate accurate and coherent summaries. This section provides an overview of datasets commonly used in ATS research, including their characteristics, domains, and the specific challenges they address. We categorize these datasets based on their application in extractive, abstract, or hybrid summarization techniques.

For example, the authors in [1] evaluated their model using the publicly available Multi-News dataset (version 1.0.0) from TensorFlow, which is widely used for benchmarking multi-document summarization models. This dataset includes news articles with human-generated summaries and is structured into training (44,972 document sets), test (5,622 document sets), and validation (5,622 document sets) splits. Another popular dataset, used in studies such as [9], [11], and [12], is the BBC News dataset, which contains around 2,225 articles across five domains: business, entertainment, politics, sports, and technology. The dataset is split into 90% for training and 10% for testing, ensuring that each genre is proportionately represented. Additionally, [6] leveraged the WikiHow dataset, consisting of instructional articles with human-generated summaries, while [7] used a diverse set of datasets: CNN/Daily Mail for news articles, PubMed for biomedical literature, and ArXiv for research papers, supporting a range of summarization needs across domains.

The DUC and TAC datasets [9], [15], [17], and [18], commonly employed for extractive summarization, provide comprehensive benchmarks. For instance, DUC-2003 and DUC-2004 are used for training and testing, while the TAC 2008–2010 datasets are useful for evaluating multi-document summarization models. Other important datasets include the Annotated Gigaword [23], which focuses on single-document summarization, and the Samsun dataset [52], which specializes in dialogue-based summarization, offering a rich resource for conversational models.

Furthermore, datasets such as the NEWS SUMMARY from Kaggle [21] and the PubMed citations from MEDLINE [22] provide varied

resources for training models in different domains. The NEWS SUMMARY dataset includes over 100,000 news articles, while the PubMed dataset offers a rich collection of biomedical abstracts. Other specialized datasets include RegNEWS [55] (news articles with transliterated words) and Project Gutenberg [24] (literary chapter-summary pairs), which provide valuable resources for domain-specific summarization research.

While significant progress has been made in text summarization for high-resource languages like English, efforts to develop datasets for low-resource languages are gaining momentum. For instance, a notable Arabic dataset for text summarization is SumArabic [61]. This dataset comprises 84,764 high-quality text-summary pairs extracted from two Arabic news websites: emaratalyoum.com and almamlakatv.com. It is divided into training (75,817 pairs), validation (4,121 pairs), testing (4,174 pairs), and out-of-domain (652 pairs) sets. SumArabic is designed for abstractive summarization tasks, providing a valuable resource for developing and evaluating summarization models in the Arabic language.

Furthermore, the LANS [62] (Large-scale Arabic News Summarization) dataset offers 8.4 million articles and their summaries, extracted from newspaper websites' metadata between 1999 and 2019. The summaries, written by journalists from 22 major Arab newspapers, cover diverse topics, enhancing the dataset's applicability for various summarization tasks.

In the context of Indian languages, [63] used the ILSUM initiative which has introduced datasets for languages such as Hindi, Gujarati, Bengali, Kannada, Tamil, and Telugu. These datasets comprise over 15,000 news articles for each language, paired with corresponding headlines, facilitating research in summarization for these languages.

For Italian, in [64] two new datasets have been developed: Fanpage and IIPost. These datasets consist of multi-sentence summaries and corresponding articles collected from Italian news websites. They serve as valuable resources for abstractive summarization in Italian, a language previously lacking in such datasets.

In conclusion, the datasets discussed here contribute significantly to the progress of text summarization research. They offer diverse characteristics and applications, aiding in the development of models tailored to various domains. A summary of all reviewed datasets, including their sizes, domains, and types of summarizations, is presented in Table 1, consolidating the key resources available for researchers.

## 7. Evaluation Methods

Early research in summary evaluation focused on human judges [31, 32]. However, due to the high cost and impracticality of manual evaluation for large datasets, researchers have developed automatic metrics to assess summary quality more efficiently. Summary evaluation is generally performed in two ways: (1) using automatic evaluation tools and (2) manual evaluation [15]. These evaluations typically measure four key dimensions: coherence, consistency, fluency, and relevance [33]. Recently, Fabbri et al. [34] conducted a meta-evaluation of commonly used automatic metrics, incorporating crowd-sourced human annotations for a more robust assessment.

### 7.1. Similarity-Based Summary Evaluation

Most summary evaluation metrics rely on the similarity between the generated summary and a human-written reference summary. ROUGE F-scores [35] have been the standard for assessing summarization model performance, measuring n-gram lexical overlap between reference and candidate summaries. ROUGE-1 and ROUGE-2 scores measure unigram and bigram overlap [21], respectively, to indicate informativeness, while ROUGE-L assesses the longest common subsequence to gauge summary fluency. However, ROUGE's reliance on exact n-gram matches overlooks semantic overlaps between synonymous phrases, penalizing models that produce novel wordings. To address this limitation, researchers have explored similar metrics based on contextualized embeddings, including BERTScore [36], MoverScore [37], and Sentence Mover's Similarity [38]. These methods embed candidate and reference summaries as vectors using pre-trained encoders, though they may introduce biases and lack interpretability. BartScore [39] approaches evaluation as a text generation problem, where models trained to convert generated text to/from a reference or the original source text achieve higher scores, reflecting better quality.

### 7.2. Factual Consistency

Factual consistency is crucial for summaries, especially in domains requiring accuracy [40, 41]. The authors in [23] investigated the correlation between factual consistency and human evaluation. Several metrics have been developed to assess the faithfulness of summaries, leveraging techniques based on text entailment or question answering (QA). In text entailment-based approaches, factual inconsistencies are detected by verifying the alignment between the summary and the original document. For instance, FactCC [41], a weakly supervised metric, uses a BERT-based model trained

on rule-based transformations of source sentences to identify factual errors. QA-based metrics, such as FEQA [42] and QAGS [43], generate questions from a given summary and assess factual consistency by verifying whether the summary contains sufficient information to answer these questions accurately. QuestEval [44] combines precision and recall-based QA metrics to evaluate factual consistency by measuring how well the summary can answer questions derived from the source content. While the measures of factual consistency focus on qualitative evaluation, ROUGE metrics provide a quantitative assessment [23].

### 7.3. Coherence and Redundancy

Coherence and redundancy are essential aspects of summary quality, ensuring clarity and conciseness. The SNaC framework [45] evaluates narrative coherence, particularly for long summaries, using fine-grained annotations. For redundancy, Peyrard et al. [46] propose a metric that measures the extent of redundancy in a summary by calculating the ratio of unique n-grams within the text. Xiao and Carenini [47] introduce an alternative approach, defining redundancy as the inverse of a diversity metric with length normalization. Here, diversity is based on the entropy of unigrams in the document.

Table.1 Datasets Employed in Text Summarization Studies

Ref.	Dataset Name	Size	Language	Domain	Single Document	Multi-Documents	Link
[1]	Multi-News	56,216	English	news articles and human-written summaries	x	✓	<a href="https://github.com/Alex-Fabbri/Multi-News">https://github.com/Alex-Fabbri/Multi-News</a>
[6]	WikiHow	230,843	English	articles about various topics (from arts and entertainment to computers and electronics)	x	✓	<a href="https://paperswithcode.com/dataset/wikihow">https://paperswithcode.com/dataset/wikihow</a>
[7]	CNN/DM	312,085	English	short news articles	x	✓	<a href="https://github.com/QianRuan/histruct/releases/tag/data_and_models">https://github.com/QianRuan/histruct/releases/tag/data_and_models</a>
	PubMed	133,215	English	biomedical literature	x	✓	
	ArXiv	215,913	English	research papers from multiple domains	x	✓	
[8]	MultiLing 2015	10,000	Multiple languages, but the work was restricted to English	Wikipedia articles	✓	x	<a href="http://multiling.iit.demokritos.gr/pages/view/1516/multiling-2015">http://multiling.iit.demokritos.gr/pages/view/1516/multiling-2015</a>
[9]	DUC 2003 and DUC 2004	DUC 2003: Contains 30 topics, each with around 10 documents per topic. DUC 2004: Contains 50 topics, also with multiple related documents per topic.	English	News articles	x	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
	TAC (2008-2011)	TAC 2008 to TAC 2011 each contain about 48 topics, with 10	English	News articles	x	✓	

		documents per topic.					
[10], [12], [14]	BBC-article news summary	2225	English	Entertainment, Sport, Politics, Business, and Technical	✓	x	<a href="https://www.kaggle.com/datasets/pariza/bbc-news-summary">https://www.kaggle.com/datasets/pariza/bbc-news-summary</a>
	DUC2005	500	English	news, including areas like politics, technology, sports, and international events.	✓	✓	
[11]	DUC2007	500	English	news, including areas like politics, technology, sports, and international events.	x	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
	DUC01	309	English	Multi-Domain	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
	DUC02	567	English	Multi-Domain	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
[13]	CNN	3000	English	Multi-Domain	✓	x	<a href="https://github.com/abisee/cnn-dailymail">https://github.com/abisee/cnn-dailymail</a>
	DUC2001	309	English	Multi-Domain	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
	DUC2002	567	English	Multi-Domain	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
[15]	Daily mail	170K	English	news, entertainment, health, and lifestyle	✓	x	<a href="https://github.com/JafferWilson/Process-Data-of-CNN-DailyMail">https://github.com/JafferWilson/Process-Data-of-CNN-DailyMail</a>
	C Irvine ML Repository	1-22 KB per file	English	Sports & News Articles	✓	x	<a href="https://archive.ics.uci.edu/">https://archive.ics.uci.edu/</a>
	DUC2004	Varies	English	News Articles	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
[16]	BBC News	2225	English	News Articles	✓	x	<a href="https://www.kaggle.com/datasets/pariza/bbc-news-summary">https://www.kaggle.com/datasets/pariza/bbc-news-summary</a>
	DUC-2002	Varies	English	News Articles	✓	✓	
	DUC-2003	Varies	English	News Articles	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
[17]	DUC-2004	Varies	English	News Articles	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
	TAC-11	Varies	English	News Articles	✓	✓	
	DUC2001	309	English	Multi-Domain	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
[18]	DUC2002	567	English	Multi-Domain	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
	CNN	85,881	English	News Articles	✓	x	<a href="https://github.com/abisee/cnn-dailymail">https://github.com/abisee/cnn-dailymail</a>
[19]	Daily mail	219,100	English	News Articles	✓	x	<a href="https://github.com/abisee/cnn-dailymail">https://github.com/abisee/cnn-dailymail</a>
[20] [51]	CNN/DailyMail	~1 million articles	English	News Articles	✓	x	<a href="https://github.com/QianRuan/histruct/releases/tag/data_and_models">https://github.com/QianRuan/histruct/releases/tag/data_and_models</a>
[21]	NEWS SUMMARY	4515 articles	English	News Articles	✓	x	<a href="https://www.kaggle.com/datasets/sunysai12345/news-summary">https://www.kaggle.com/datasets/sunysai12345/news-summary</a>

[22]	PubMed (Subset)	14 million citations	English	biomedical literature	✓	x	<a href="https://github.com/QianRuan/histruct/releases/tag/data_and_models">https://github.com/QianRuan/histruct/releases/tag/data_and_models</a>
[23]	Annotated Gigaword	3.8 million pairs	English	News Articles	✓	x	<a href="https://huggingface.co/datasets/Harvard/gigaword">https://huggingface.co/datasets/Harvard/gigaword</a>
	CNN/DailyMail	287,227 pairs	English	News Articles	x	✓	<a href="https://github.com/QianRuan/histruct/releases/tag/data_and_models">https://github.com/QianRuan/histruct/releases/tag/data_and_models</a>
[24]	Novel summary-chapter pairs	8088 pairs	English	Literature/Education	✓	x	<a href="https://github.com/manestay/novel-chapter-dataset">https://github.com/manestay/novel-chapter-dataset</a>
[52]	SAMSum	~ 16,000 dialogues	English	conversational text	✓	x	<a href="https://www.kaggle.com/datasets/nileshmalode1/samsu-m-dataset-text-summarization">https://www.kaggle.com/datasets/nileshmalode1/samsu-m-dataset-text-summarization</a>
[53]	Chinese News,	1200	Chinese	News texts	✓	x	N/A
	DUC-2004	Varies	English	News Articles	✓	✓	<a href="https://github.com/danieldeutsch/duc-tac-data">https://github.com/danieldeutsch/duc-tac-data</a>
[54]	Wikipedia Content, PDF Documents, Plain Text	Varies	English	N/A	✓	✓	N/A
[55]	NEWS SUMMARY	102,915	English	News Articles, Hindu, Times of India, the Guardian, and various other sources along with human-generated abstractive summaries	✓	x	<a href="https://www.kaggle.com/datasets/sunysai12345/news-summary">https://www.kaggle.com/datasets/sunysai12345/news-summary</a>
	RegNEWS	50,246	English	News, transliterated words	✓	x	N/A
[61]	SumArabic	84,764 pairs	Arabic	News articles	✓	✓	<a href="https://data.mendeley.com/datasets/7kr75c9h24/1">https://data.mendeley.com/datasets/7kr75c9h24/1</a>
[62]	LANS	8.4 million articles	Arabic	New articles	✓	✓	N/A
[63]	ILSUM	15,000 articles /language	Hindi, Gujarati, Bengali, Kannada, Tamil, Telugu	New	✓	✓	<a href="https://ilsum.github.io/ilsum/2024/?utm_source=chatgpt.com">https://ilsum.github.io/ilsum/2024/?utm_source=chatgpt.com</a>
[64]	Fanpage, IIPost	Varies (Thousands of pairs)	Italian	New	✓	✓	N/A

## 8. Literature Review

Several studies have proposed extractive summarization methods that aim to identify and select important sentences or segments directly from the input text. These methods often rely on features such as term frequency, sentence position, and length. For example, in [1], a statistical approach is used to select key sentences based on these features to capture essential information. Similarly, Nikolaos G. et al. [8] propose an approach combining binary classification and topic-based sentence extraction using Latent Dirichlet Allocation (LDA). In contrast, [6] introduces a method that clusters document content based on latent topics, summarizing each cluster to form a cohesive summary, with promising results on the WikiHow dataset.

A more advanced extractive model is proposed in [7], which incorporates local topic extraction and hierarchical modules to overcome transformer limitations, significantly improving summary accuracy as evaluated on large datasets. Additionally, [10] develops a graph-based summarization method, where sentences are modeled as graph nodes, improving precision and recall on the BBC news dataset.

Other studies focus on improving summary diversity and reducing redundancy. For instance, [15] presents a model that learns to select sentences based on both saliency and dissimilarity, outperforming traditional methods on the DUC and Daily Mail datasets. Furthermore, fuzzy logic techniques have been explored in [16], enhancing reliability and clarity by combining it with extractive summarization methods. A related work by [17] proposes an extractive summarization model using Shark Smell Optimization (SSO) and fuzzy logic, evaluated on various datasets and showing improved ROUGE scores.

**Abstractive Summarization:** Abstractive summarization methods, which generate summaries by paraphrasing the original text, have also seen significant advancements. [13] introduces two models, EXABSUM extractive and EXABSUM abstractive, where the latter uses a word graph and keyphrase reranking to generate abstractive summaries. Their approach outperforms traditional extractive methods and competes with state-of-the-art abstractive models on multi-domain benchmarks. Additionally, deep learning-based approaches for abstractive summarization have gained traction. [22] proposes a reinforcement learning-based model for biomedical publications, generating domain-aware summaries with metrics based on biomedical expert tools. Similarly, [19] integrates a graph-based attention mechanism with sequence-to-sequence frameworks, addressing the saliency factor that prior

models overlooked, and improving results over earlier neural abstractive methods.

Furthermore, hybrid approaches combining extractive and abstractive methods have also been explored. In [24], a pipeline approach is introduced, where an extractive component filters content for the abstractive part, achieving improved ROUGE-1 scores. This model is further refined by a margin ranking loss to separate positive and negative examples during extraction.

**Hybrid Models and New Directions:** The potential of hybrid summarization models is demonstrated by various studies. For example, [51] presents a hybrid system that integrates deep contextualized embeddings with statistical features, and [52] proposes a framework combining extractive and abstractive components, demonstrating improvements in coherence and informativeness. Liu et al. [53] similarly integrate both paradigms within a single framework, reducing redundancy and enhancing accuracy.

In the domain of multilingual and transliterated text, Muniraj et al. [55] introduce HNTSumm, a model designed to handle transliterated news articles, improving summarization accuracy in multilingual contexts. This highlights the adaptability of hybrid models in handling complex document structures across languages.

Transformers architectures and Large Language Models (LLMs) have significantly advanced the field of text summarization, offering enhanced capabilities in processing and generating human-like summaries.

In [65], Traditional Transformer models face challenges with long sequences due to their quadratic scaling of self-attention mechanisms. To address this, Longformer introduces an attention mechanism that scales linearly with sequence length, enabling efficient processing of lengthy documents. Its performance on tasks like WikiHop and TriviaQA demonstrates its effectiveness in handling long inputs.

Similarly [66], LongT5 integrates attention mechanisms from long-input Transformers and pre-training strategies from summarization models, achieving state-of-the-art results in summarization tasks and outperforming original T5 models in question-answering tasks.

The study [67] explores architectural modifications and pretraining strategies to adapt Transformers for long inputs. The proposed staggered, block-local Transformer with global encoder tokens, along with additional pretraining on long sequences, enhances summarization performance without extensive increases in model size. This approach led to the development of PEGASUS-X, capable of handling inputs up to 16K tokens.

A comparative study [68] analyzes models such as MPT-7b-instruct, Falcon-7b-instruct, and OpenAI's

ChatGPT text-davinci-003. Evaluations using metrics like BLEU, ROUGE, and BERT Scores indicate that text-davinci-003 outperforms the others in generating high-quality summaries.

Also [69], Multimodal Summarization Addresses the integration of multimedia in scientific publications, the Uni-SciSum framework combines text, images, video, and audio to produce multimodal summaries. Leveraging LLMs and a query-based Transformer, BridgeNet fuses diverse modalities, enhancing the richness and informativeness of summaries.

Domain-specific summarization refers to the creation of summaries tailored to a specific field, industry, or subject matter. Unlike generic summarization, which focuses on general linguistic features, domain-specific summarization needs to incorporate specialized knowledge, terminology, and nuances that are unique to a particular domain. The necessity for domain-specific models arises from the fact that certain fields such as legal, medical, financial, and scientific domains use domain-specific language and conventions that general-purpose summarization models often fail to capture effectively. By leveraging domain-specific datasets, specialized knowledge, and fine-tuning techniques, domain-specific summarization aims to produce summaries that are not only concise but also accurate and contextually relevant.

In the legal domain, documents such as case law, contracts, and regulations contain complex language, technical terminology, and long sentences. Traditional summarization methods, which rely on general linguistic features, often fail to capture the subtleties of legal language. Recent approaches have therefore focused on fine-tuning transformer models, such as BERT and T5, to create more accurate legal summaries.

In [70], pre-trained Transformer models, while dominant in most NLP tasks, face input length limitations in the legal domain. Even sparse-attention models like Longformer and BigBird struggle with truncating long legal texts. The study explores two solutions: (i) extending Longformer with LegalBERT to handle longer texts (up to 8,192 sub-words), and (ii) adapting LegalBERT to use TF-IDF representations. The first approach outperforms previous models, while the second offers more efficiency but lower performance, still outperforming linear SVM with TF-IDF for long legal document classification.

Financial documents, such as annual reports, earnings calls, and market analyses, often contain dense and highly specialized content. Summarizing such documents requires understanding financial terms, relationships between entities (e.g., stocks, companies), and domain-specific metrics. The complexity of financial documents makes traditional summarization methods unsuitable for producing summaries that are both accurate and contextually relevant.

In [71], the authors describe their participation in the FinNLP-AgentScen 2024 shared task #2 on Financial Text Summarization, where they fine-tuned a foundation model for the finance domain. Their approach involved (1) adapting the Llama3 8B model to finance through continued pre-training, (2) applying multi-task instruction-tuning to enhance the model's finance-related capabilities, and (3) fine-tuning it to become a task-specific model. Their model, FinLlama3\_sum, achieved strong performance, securing third place in its category with a ROUGE-1 score of 0.521.

Furthermore, [72] the authors introduce FIN2SUM, a framework designed for summarizing the managerial analysis and discussion sections of 10-K reports from top NASDAQ-listed companies. The research focuses on evaluating Large Language Models (LLMs) for financial text summarization, with a particular emphasis on LLAMA-2's capability in handling complex financial content. The study compares three state-of-the-art LLMs, LLAMA-2, FLAN, and Claude 2 using BERT and ROUGE scores for evaluation. The results demonstrate that FIN2SUM, powered by LLAMA-2, significantly enhances AI-driven financial text summarization, offering a valuable tool for analysts and decision-makers.

In conclusion, the literature on text summarization has evolved significantly, with both extractive and abstractive methods contributing to the development of more accurate and efficient models. While traditional approaches have relied heavily on statistical methods, clustering, and graph-based techniques, recent advancements in Transformers-based architecture and Large Language Models (LLMs) have pushed the boundaries of summarization performance. These models, including BERT, T5, and GPT-based approaches, have demonstrated remarkable abilities in generating high-quality summaries, addressing challenges like redundancy, domain adaptation, and long document handling. Additionally, multimodal summarization techniques are gaining traction, further expanding the possibilities of summarization beyond text alone. Also, Domain-specific summarization models are critical for producing meaningful, contextually aware summaries of documents in specialized fields. By fine-tuning large pre-trained models such as BERT and T5 on domain-specific data, these models are able to better understand the unique language and structure of their respective fields, leading to more accurate and informative summaries. The advances in legal, financial, and medical summarization, particularly through the use of hybrid models, knowledge graphs, and multi-task learning, have significantly improved the quality of summaries generated for these specialized domains. Despite impressive progress, challenges remain, such as the scalability of models for long inputs, the

management of factual consistency, and the trade-offs between model complexity and performance.

## 9. Challenges and Future Directions

Automatic text summarization techniques face several challenges, including dealing with the complexity of context understanding, which remains a hurdle for both extractive and abstractive methods. One key challenge is the difficulty in handling long documents, where both local and global context must be effectively integrated for accurate summaries [1][7]. The ability to maintain coherence and factual consistency, particularly in abstractive summarization, is another ongoing issue, as models often generate summaries that may be fluent but inaccurate or incomplete in terms of information representation [19][40]. Moreover, domain-specific summarization faces the challenge of tailoring models to handle technical or highly specialized content, which may require significant adjustments to improve performance in these fields [50][28]. Another challenge is the redundancy present in lengthy documents, which can lead to repetitive summaries that fail to provide concise information [47]. Furthermore, the evaluation of summarization models remains problematic due to the lack of universally accepted metrics that can effectively measure content relevance and summary quality [34][46]. Additionally, despite advances, extractive methods still struggle with the extraction of meaningful content, especially when dealing with complex sentence structures or nuanced language [11][10]. Finally, hybrid models that combine extractive and abstractive methods aim to address these challenges but face difficulties in balancing the strengths of both approaches, as the integration of the two requires careful alignment to avoid generating incoherent or disjointed summaries [52][53].

Furthermore, Transformer-based and LLM-based approaches dominate current research, offering flexible, high-quality, and controllable summarization capabilities. However, challenges remain, including factual consistency, hallucination, and domain adaptation, motivating ongoing research in fine-tuning, evaluation, and controlled generation. Recent advances in transformer-based models and LLMs provide promising directions for future research, and ongoing efforts are focused on addressing the limitations posed by these models, especially in the context of domain-specific and large-scale text summarization.

In the era of Large Language Models [68] such as GPT-4, Claude, and MPT, new challenges have emerged that compound or transform older limitations. A major concern is factual consistency. LLMs are prone to hallucinations, generating information that sounds plausible but is factually incorrect, which undermines their reliability for real-

world applications. Prompt sensitivity is another issue, where slight changes in user input can lead to substantially different outputs, making it difficult to ensure reproducibility and control in summarization tasks. Legal and ethical considerations, such as unintentional plagiarism, content bias, and privacy risks, also complicate the deployment of LLMs for summarization, particularly in regulated domains like healthcare or finance. Furthermore, LLMs remain largely black-box systems, making explainability a persistent challenge, especially when understanding how summaries are derived from the input. High computational costs and latency further limit the real-time usability of LLM-based summarization, particularly for large-scale or interactive applications. Despite these limitations, LLMs offer flexible, high-quality, and controllable summarization capabilities, making them central to ongoing efforts that aim to mitigate these shortcomings through techniques such as reinforcement learning from human feedback (RLHF), improved evaluation metrics (e.g., QAFactEval, FactScore), and structured prompting frameworks. As the field progresses, addressing these LLM-specific limitations while learning from traditional challenges such as data sparsity and domain adaptation will be critical for developing robust, trustworthy, and scalable summarization systems.

Looking ahead, future directions in text summarization could focus on improving the understanding of context, enhancing the ability to generate summaries that reflect deep semantic knowledge, and addressing limitations in handling long and complex documents. Incorporating multimodal information, such as visual and auditory content alongside text, could further enrich summaries, making them more comprehensive and context aware. Furthermore, research into more robust and interpretable evaluation metrics will be crucial in assessing the true quality of generated summaries, especially when dealing with diverse languages and domains. Another promising direction involves the use of reinforcement learning to continuously improve summarization models, as well as increasing the focus on real-time summarization for applications like news aggregation and live content monitoring. Overall, the future of text summarization holds great potential for both advancing the underlying technologies and expanding their practical applications across various industries.

## 10. Conclusion

In conclusion, Text summarization has made significant progress in recent years, driven by advancements in machine learning, natural language processing, and, more recently, large language models. Traditional extractive and abstractive

methods have offered unique strengths. Extractive approaches being more reliable but prone to redundancy, and abstractive ones capable of generating fluent language but often struggle with factual consistency and deep context understanding. Hybrid models have emerged to bridge these approaches, yet effectively balancing their strengths remains a complex task. In this survey, we introduced a novel multi-dimensional taxonomy that reflects the latest trends in summarization, including LLM-based, domain-specific, and multimodal approaches. Instruction-tuned LLMs such as GPT-4, Claude, Falcon, and MPT now enable powerful zero-shot and customizable summarization, marking a shift from dataset-specific training toward more generalized and controllable systems. However, these models bring new challenges including hallucinations, prompt sensitivity, legal and ethical concerns, explainability, and high computational costs that compound earlier limitations like data sparsity and poor domain adaptation. Furthermore, reliable evaluation remains an open problem. While classical metrics like ROUGE and BLEU persist, more newer metrics are increasingly necessary to assess the factual and contextual quality of summaries generated by LLMs. Despite these issues, the continued evolution of summarization models offers great promise for applications in content creation, information retrieval, scientific writing, and real-time summarization. Looking ahead, addressing the unique limitations of LLMs while building on the strengths of earlier methods will be key to developing robust, trustworthy, and scalable summarization systems.

**Competing Interests:** The authors declare no competing interests.

#### Authors' contribution:

**Sara Zayed:** Conceptualization, Methodology, Software, Writing the original draft, Writing review & editing. **Mostafa Ezzat:** Supervision, Data curation, Validation, Visualization, Investigation, Writing review & editing. **Hesham A. Hefny:** Supervision, Data curation, Visualization, Investigation, review & editing. All authors reviewed the results and approved the final version of the manuscript.

#### Declaration of Generative AI and AI-assisted Technologies in the Writing Process:

During the preparation of this work, the authors used OpenAI's ChatGPT4 to refine text quality. After using this tool, the authors reviewed and edited the content needed and took full responsibility for the content of the publication.

## References

- [1] K. Tewari, A. K. Yadav, M. Kumar, and D. Yadav, "Extractive text summarization using statistical approach," in *Computer Vision and Machine Intelligence*, M. Tistarelli, S. R. Dubey, S. K. Singh, and X. Jiang, Eds. Singapore: Springer, 2023, pp. 633–643, doi: 10.1007/978-981-19-7867-8\_52.
- [2] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, p. 113679, 2021, doi: 10.1016/j.eswa.2020.113679.
- [3] G. C. V. Vilca and M. A. S. Cabezudo, "A study of abstractive summarization using semantic representations and discourse level information," in *Proc. 20th Int. Conf. Text, Speech, Dialogue*, Prague, Czech Republic, 2017.
- [4] V. Patel and N. Tabrizi, "An automatic text summarization: A systematic review," *Computación y Sistemas*, vol. 26, no. 3, 2022, doi: 10.13053/cys-26-3-4347.
- [5] I. Mani and M. T. Maybury, Eds., *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.
- [6] K. A. R. Issam, S. Patel, and S. C. N., "Topic modeling based extractive text summarization," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 6, pp. 1710–1719, 2020, doi: 10.35940/ijitee.F4611.049620.
- [7] T. Wang et al., "A study of extractive summarization of long documents incorporating local topic and hierarchical information," *Sci. Rep.*, vol. 14, p. 10140, 2024, doi: 10.1038/s41598-024-60779-z.
- [8] N. Gialitsis, N. Pittaras, and P. Stamatopoulos, "A topic-based sentence representation for extractive text summarization," in *Proc. Workshop MultiLing 2019: Summarization Across Languages, Genres and Sources*, Varna, Bulgaria, 2019, pp. 26–34.
- [9] O. Ernst et al., "Proposition-level clustering for multi-document summarization," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Seattle, WA, USA, 2022, pp. 1765–1779.
- [10] A. K. Yadav, Ranvijay, R. S. Yadav et al., "Graph-based extractive text summarization based on single document," *Multimed. Tools Appl.*, vol. 83, pp. 18987–19013, 2024, doi: 10.1007/s11042-023-16199-8.
- [11] Waseemullah et al., "A novel approach for semantic extractive text summarization," *Appl. Sci.*, vol. 12, no. 9, p. 4479, 2022, doi: 10.3390/app12094479.
- [12] S. Mandal et al., "Extractive text summarization using supervised learning and natural language processing," in *Proc. 2021 Int. Conf. Intell. Technol. (CONIT)*, Hubli, India, 2021, pp. 1–7, doi: 10.1109/CONIT51480.2021.9498322.
- [13] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "EXABSUM: A new text summarization approach for generating extractive and abstractive summaries," *J. Big Data*, vol. 10, p. 163, 2023, doi: 10.1186/s40537-023-00836-y.

- [14] A. K. Yadav et al., “Extractive text summarization using deep learning approach,” *Int. J. Inf. Technol.*, vol. 14, pp. 2407–2415, 2022, doi: 10.1007/s41870-022-00863-7.
- [15] S. R. Chowdhury and K. Sarkar, “A new method for extractive text summarization using neural networks,” *SN Comput. Sci.*, vol. 4, p. 384, 2023, doi: 10.1007/s42979-023-01806-0.
- [16] B. Sharma et al., “Automatic text summarization using fuzzy extraction,” in *Int. Conf. Innov. Comput. Commun.*, Singapore: Springer, 2020, pp. 411–422, doi: 10.1007/978-981-15-1286-5\_33.
- [17] M. Tomer et al., “Enhancing metaheuristic based extractive text summarization with fuzzy logic,” *Neural Comput. Appl.*, vol. 35, pp. 9711–9723, 2023, doi: 10.1007/s00521-023-08209-5.
- [18] D. Debnath, R. Das, and P. Pakray, “Single document text summarization addressed with a cat swarm optimization approach,” *Appl. Intell.*, vol. 53, pp. 12268–12287, 2023, doi: 10.1007/s10489-022-04149-0.
- [19] J. Tan, X. Wan, and J. Xiao, “Abstractive document summarization with a graph-based attentional neural model,” in *Proc. 55th Annu. Meet. Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, Vancouver, Canada, 2017, pp. 1171–1181.
- [20] T. Yu and S. Gao, “Abstractive text summarization with semantic dependency graph,” in *Proc. 2023 5th Int. Acad. Exchange Conf. Sci. Technol. Innov. (IAECST)*, Guangzhou, China, 2023, pp. 567–570, doi: 10.1109/IAECST60924.2023.10502964.
- [21] G. Karuna et al., “Automated abstractive text summarization using deep learning,” *E3S Web Conf.*, vol. 430, p. 01021, 2023, doi: 10.1051/e3sconf/202343001021.
- [22] P. Gigioli et al., “Domain-aware abstractive text summarization for medical documents,” in *Proc. 2018 IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, Madrid, Spain, 2018, pp. 2338–2343, doi: 10.1109/BIBM.2018.8621539.
- [23] P. Kouris, G. Alexandridis, and A. Stafylopatis, “Text summarization based on semantic graphs: An abstract meaning representation graph-to-text deep learning approach,” *J. Big Data*, vol. 11, p. 95, 2024, doi: 10.1186/s40537-024-00950-5.
- [24] H. Hardy et al., “Novel chapter abstractive summarization using spinal tree aware sub-sentential content selection,” arXiv:2211.04903, 2022, doi: 10.48550/arXiv.2211.04903.
- [25] K. Shyamala and M. M. Evangeline, “A framework for extractive text summarization of single text document in Tamil language using frequency based feature extraction technique,” in *Lect. Notes Netw. Syst.*, 2023, pp. 155–164, doi: 10.1007/978-981-99-6755-1\_12.
- [26] W. Zhang, Z. Xue, and L. Zhao, “Recent advances in tokenization methods for natural language processing,” *Comput. Linguist. J.*, vol. 11, no. 5, pp. 357–370, 2021.
- [27] Y. Dong, “A survey on neural network-based summarization methods,” arXiv:1804.04589, 2018, doi: 10.48550/arXiv.1804.04589.
- [28] N. Supriyono et al., “A survey of text summarization: Techniques, evaluation and challenges,” *Nat. Lang. Process. J.*, vol. 7, p. 100070, 2024, doi: 10.1016/j.nlp.2024.100070.
- [29] G. A. M. Mendoza, Y. Ledeneva, and R. A. García-Hernández, “Determining the importance of sentence position for automatic text summarization,” in *Proc. 2020 Int. Conf. Comput. Sci. Appl.*, 2020, pp. 2421–2431.
- [30] M. Jain and H. Rastogi, “Automatic text summarization using soft-cosine similarity and centrality measures,” in *Proc. 2020 4th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Coimbatore, India, 2020, pp. 1021–1028, doi: 10.1109/ICECA49313.2020.9297583.
- [31] M. Cao, “A survey on neural abstractive summarization methods and factual consistency of summarization,” arXiv:2204.09519, 2022, doi: 10.48550/arXiv.2204.09519.
- [32] A. Nenkova and R. J. Passonneau, “Evaluating content selection in summarization: The pyramid method,” in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics (HLT-NAACL 2004)*, 2004, pp. 145–152.
- [33] M. Zhong et al., “Towards a unified multi-dimensional evaluator for text generation,” in *Proc. 2022 Conf. Empir. Methods Nat. Lang. Process.*, 2022, pp. 2023–2038.
- [34] A. R. Fabbri et al., “SummEval: Re-evaluating summarization evaluation,” *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 391–409, 2021.
- [35] C.-Y. Lin and E. Hovy, “Automatic evaluation of summaries using n-gram co-occurrence statistics,” in *Proc. 2003 Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2003, pp. 150–157.
- [36] T. Zhang et al., “BERTScore: Evaluating text generation with BERT,” in *Int. Conf. Learn. Represent.*, 2020.
- [37] W. Zhao et al., “MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance,” arXiv:1909.02622, 2019.
- [38] E. Clark, A. Celikyilmaz, and N. A. Smith, “Sentence mover’s similarity: Automatic evaluation for multi-sentence texts,” in *Proc. 57th Annu. Meet. Assoc. Comput. Linguistics*, 2019, pp. 2748–2760.
- [39] W. Yuan, G. Neubig, and P. Liu, “BARTScore: Evaluating generated text as text generation,” *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 27263–27277, 2021.
- [40] J. Maynez et al., “On faithfulness and factuality in abstractive summarization,” arXiv:2005.00661, 2020.
- [41] W. Kryscinski, B. McCann, C. Xiong, and D. Socher, “Evaluating the factual consistency of abstractive text summarization,” in *\*Proc. 2020 Conf. Empir.*

- Methods Nat. Lang. Process. (EMNLP)\*, 2020, pp. 9332–9346, doi: 10.18653/v1/2020.emnlp-main.750.
- [42] E. Durmus, H. He, and M. Diab, “FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization,” \*arXiv preprint arXiv:2005.03754\*, 2020.
- [43] A. Wang, K. Cho, and M. Lewis, “Asking and answering questions to evaluate the factual consistency of summaries,” \*arXiv preprint arXiv:2004.04228\*, 2020.
- [44] T. Scialom et al., “QuestEval: Summarization asks for fact-based evaluation,” in \*Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.\*, 2021, pp. 6594–6604.
- [45] T. Goyal, J. J. Li, and G. Durrett, “SNaC: Coherence error detection for narrative summarization,” in \*Proc. 2022 Conf. Empir. Methods Nat. Lang. Process.\*, 2022, pp. 444–463.
- [46] M. Peyrard, T. Botschen, and I. Gurevych, “Learning to score system summaries for better content selection evaluation,” in \*Proc. Workshop New Frontiers Summarization\*, 2017, pp. 74–84.
- [47] W. Xiao and G. Carenini, “Systematically exploring redundancy reduction in summarizing long documents,” in \*Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics Int. Joint Conf. Nat. Lang. Process.\*, 2020, pp. 516–528.
- [48] G. R. Sundar et al., “Machine learning based extractive text summarization techniques,” in \*Proc. 2023 Int. Conf. Integr. Intell. Commun. Syst. (ICIICS)\*, Kalaburagi, India, 2023, pp. 1–6, doi: 10.1109/ICIICS59993.2023.10421192.
- [49] C. Badgujar, V. Jethani, and T. Ghorpade, “Abstractive summarization using graph based methods,” in \*Proc. 2018 2nd Int. Conf. Invent. Commun. Comput. Technol. (ICICCT)\*, Coimbatore, India, 2018, pp. 803–807, doi: 10.1109/ICICCT.2018.8473315.
- [50] A. Afzal et al., “Challenges in domain-specific abstractive summarization and how to overcome them,” in \*Proc. 14th Int. Conf. Agents Artif. Intell.\*, 2023, pp. 682–689, doi: 10.5220/0011744500003393.
- [51] M. Gambhir and V. Gupta, “Improved hybrid text summarization system using deep contextualized embeddings and statistical features,” \*Multimed. Tools Appl.\*, 2024, doi: 10.1007/s11042-024-19524-x.
- [52] R. Habu et al., “A hybrid extractive-abstractive framework with pre & post-processing techniques to enhance text summarization,” in \*Proc. 2023 13th Int. Conf. Adv. Comput. Inf. Technol. (ACIT)\*, Wrocław, Poland, 2023, pp. 529–533, doi: 10.1109/ACIT58437.2023.10275584.
- [53] W. Liu et al., “A combined extractive with abstractive model for summarization,” \*IEEE Access\*, vol. 9, pp. 43970–43980, 2021, doi: 10.1109/ACCESS.2021.3066484.
- [54] J. M. Karanja and A. Matheka, “A hybrid model for text summarization using natural language processing,” \*Open J. Inf. Technol.\*, vol. 5, no. 2, pp. 65–80, 2022, doi: 10.32591/coas.ojit.0502.03065k.
- [55] P. Muniraj et al., “HNTSumm: Hybrid text summarization of transliterated news articles,” \*Int. J. Intell. Netw.\*, vol. 4, pp. 53–61, 2023, doi: 10.1016/j.ijin.2023.03.001.
- [56] M. M. Abdelsamie, S. S. Azab, and H. A. Hefny, “A comprehensive review on Arabic offensive language and hate speech detection on social media: Methods, challenges and solutions,” \*Soc. Netw. Anal. Min.\*, vol. 14, p. 111, 2024, doi: 10.1007/s13278-024-01258-1.
- [57] M. Luo, B. Xue, and B. Niu, “A comprehensive survey for automatic text summarization: Techniques, approaches and perspectives,” \*Neurocomputing\*, vol. 603, p. 128280, 2024, doi: 10.1016/j.neucom.2024.128280.
- [58] W. S. El-Kassas et al., “Automatic text summarization: A comprehensive survey,” \*Expert Syst. Appl.\*, vol. 165, p. 113679, 2021, doi: 10.1016/j.eswa.2020.113679.
- [59] G. Sharma and D. Sharma, “Automatic text summarization methods: A comprehensive review,” \*SN Comput. Sci.\*, vol. 4, p. 33, 2023, doi: 10.1007/s42979-022-01446-w.
- [60] A. P. Widyassari et al., “Review of automatic text summarization techniques & methods,” \*J. King Saud Univ. - Comput. Inf. Sci.\*, vol. 34, no. 4, pp. 1029–1046, 2022, doi: 10.1016/j.jksuci.2020.05.006.
- [61] M. B. Almarjeh, “SumArabic,” \*Mendeley Data\*, vol. 1, 2022, doi: 10.17632/7kr75c9h24.1.
- [62] A. Alhamadani et al., “LANS: Large-scale Arabic news summarization corpus,” in \*Proc. ArabicNLP 2023\*, Singapore, 2023, pp. 89–100.
- [63] A. Urlana et al., “Indian language summarization using pretrained sequence-to-sequence models,” \*arXiv preprint arXiv:2303.14461\*, 2023.
- [64] M. Kahla, Z. G. Yang, and A. Novák, “Cross-lingual fine-tuning for abstractive Arabic text summarization,” in \*Proc. Int. Conf. Recent Adv. Nat. Lang. Process. (RANLP 2021)\*, 2021, pp. 655–663.
- [65] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” \*arXiv preprint arXiv:2004.05150\*, 2020.
- [66] M. Guo et al., “LongT5: Efficient text-to-text transformer for long sequences,” \*arXiv preprint arXiv:2112.07916\*, 2021.
- [67] J. Phang, Y. Zhao, and P. J. Liu, “Investigating efficiently extending transformers for long input summarization,” \*arXiv preprint arXiv:2208.04347\*, 2022.
- [68] L. Basyal and M. Sanghvi, “Text summarization using large language models: A comparative study of MPT-7B-Instruct, Falcon-7b-Instruct, and OpenAI Chat-GPT models,” \*arXiv preprint arXiv:2310.10449\*, 2023.

- [69] Z. Tan et al., “Enhancing large language models for scientific multimodal summarization with multimodal output,” in \*Proc. 31st Int. Conf. Comput. Linguistics: Industry Track\*, Abu Dhabi, UAE, 2025, pp. 263–275.
- [70] D. Mamakas et al., “Processing long legal documents with pre-trained transformers: Modding LegalBERT and Longformer,” \*arXiv preprint arXiv:2211.00974\*, 2022.
- [71] M. Lee and S. Lay-Ki, “Finance Wizard at the FINLLM challenge task: Financial text summarization,” \*arXiv preprint arXiv:2408.03762\*, 2024.
- [72] E. Wilson et al., “FIN2SUM: Advancing AI-driven financial text summarization with LLMs,” in \*Proc. 2024 Int. Conf. Trends Quantum Comput. Emerg. Bus. Technol.\*, Pune, India, 2024, pp. 1–5, doi: 10.1109/TQCEBT59414.2024.10545078.
- [73] M. Parveen, A. Parveen, and M. M. Farooqi, “A review on automatic text summarization approaches,” \*Int. J. Comput. Appl.\*, vol. 975, pp. 8887, 2015.
- [74] S. M. A. Shah and S. A. Manan, “A survey of natural language processing techniques for summarization of unstructured text,” \*Arab. J. Sci. Eng.\*, vol. 48, pp. 7063–7083, 2023, doi: 10.1007/s13369-023-07874-7.
- [75] N. I. Kurniawan, M. R. Putra, and M. A. E. Wibawa, “Automatic text summarization using hybrid approach: A review,” \*Procedia Comput. Sci.\*, vol. 179, pp. 770–777, 2021, doi: 10.1016/j.procs.2021.01.065.
- [76] M. A. El-Haj et al., “Multilingual news summarisation: Dataset, baselines, models and metrics,” in \*Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (Volume 1: Long Papers)\*, Dublin, Ireland, 2022, pp. 3817–3833.
- [77] A. Gharebagh, A. Khadivi, and A. Jelodar, “A novel hybrid deep learning model for automatic text summarization using enhanced attention mechanism,” \*Expert Syst. Appl.\*, vol. 229, p. 120269, 2023, doi: 10.1016/j.eswa.2023.120269.
- [78] L. Chen and J. Zhuge, “A hybrid approach for abstractive summarization with sentence rewriting and entity retrieval,” in \*Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.\*, 2021, pp. 594–603.
- [79] Y. Gao et al., “SimCLS: A simple framework for contrastive learning of abstractive summarization,” in \*Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (Volume 1: Long Papers)\*, Dublin, Ireland, 2022, pp. 5565–5580.
- [80] P. Kedzie, K. McKeown, and H. Daumé III, “Content selection in deep learning models of summarization,” in \*Proc. 2018 Conf. Empir. Methods Nat. Lang. Process.\*, 2018, pp. 1818–1828.
- [81] J. Xu and G. Durrett, “Neural extractive text summarization with syntactic compression,” in \*Proc. 2019 Conf. Empir. Methods Nat. Lang. Process.\*, 2019, pp. 3292–3303.
- [82] A. Celikyilmaz, E. Hovy, and J. Gao, “Fact-based abstractive summarization,” in \*Proc. 58th Annu. Meet. Assoc. Comput. Linguist. (Volume 1: Long Papers)\*, 2020, pp. 5101–5114.
- [83] S. Narayan et al., “QURIOUS: Question generation for reading comprehension,” in \*Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.\*, 2021, pp. 7841–7857.
- [84] A. Vaswani et al., “Attention is all you need,” in \*Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS)\*, 2017, pp. 5998–6008.
- [85] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in \*Proc. 2019 Conf. Empir. Methods Nat. Lang. Process.\*, 2019, pp. 3730–3740.
- [86] A. Bhandari, A. Prabhume, and A. Black, “Re-evaluating evaluation in text summarization,” in \*Proc. 2020 Conf. Empir. Methods Nat. Lang. Process.\*, 2020, pp. 9347–9359.
- [87] A. Elgohary, M. Khader, and A. Elgohary, “Summarization of Arabic texts: A survey,” \*Egypt. Inform. J.\*, vol. 24, no. 2, pp. 153–165, 2023, doi: 10.1016/j.eij.2022.03.005.
- [88] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in \*Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist. (NAACL-HLT)\*, 2019, pp. 4171–4186.
- [89] T. Wolf et al., “Transformers: State-of-the-art natural language processing,” in \*Proc. 2020 Conf. Empir. Methods Nat. Lang. Process.: Syst. Demonstrations\*, 2020, pp. 38–45.
- [90] A. Lin et al., “DESC: A framework for domain-specific summarization evaluation,” in \*Proc. 2023 Conf. Empir. Methods Nat. Lang. Process.\*, 2023, pp. 10270–10285.
- [91] H. Ji et al., “Survey of hallucination in natural language generation,” \*ACM Comput. Surv.\*, vol. 55, no. 12, pp. 1–38, 2023, doi: 10.1145/3571730.

**Citation:** S. Zayed, M. Ezzat and H. A. Hefny. *Automatic Text Summarization: A Review of Approaches, Challenges, and Future Directions*. Journal of Computer Science & Technology, vol. 25, no. 2, pp. 87-106, 2025.

**DOI:** 10.24215/16666038.25.e08

**Received:** January 17, 2025, **Accepted:** June 13, 2025.

**Copyright:** This article is distributed under the terms of the Creative Commons License CC-BY-NC-SA.