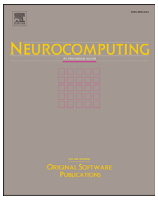
Neurocomputing 655 (2025) 131287

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey paper

A systematic review of long document summarization methods: Evaluation metrics and approaches

Bady Gana [a], Héctor Allende-Cid [a, b, c, *], Stefan Rüping [b], Marcelo Becerra-Rozas [a], Juan Zamora [d]

[a] Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Avenida Brasil 2241, Valparaíso, 2362807, Chile
[b] Knowledge Discovery, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Schloss Birlinghoven 1, Sankt Augustin, 53757, Germany
[c] Lamarr Institute for Machine Learning and Artificial Intelligence, Dortmund, 53115, Germany
[d] Instituto de Estadística, Pontificia Universidad Católica de Valparaíso, Av.Brasil 2830, Valparaíso, 2362807, Chile

HIGHLIGHTS

- Identify and categorize existing methods used for long document summarization.
- Examine how these methods are evaluated across different domains.
- Analyze the current landscape of evaluation metrics and sophisticated approaches.
- Insights for applying summarization and evaluation in large research databases.

ARTICLE INFO

Communicated by P. Li

Keywords:
Long document summarization
SLR
Text summarization evaluation
Multi-document summarization
NLP
Evaluation metrics
Deeplearning

ABSTRACT

The rapid growth of complex textual data in domains such as medicine, law, and science has heightened the relevance of Long Document Summarization (LDS). Effective summarization not only requires advanced techniques but also robust evaluation metrics capable of capturing summary quality, coherence, and factual accuracy. We analyze 113 peer-reviewed studies from last two years, selected through comprehensive searches in SCOPUS, Web of Science, and PubMed, following PRISMA 2020 guidelines. We focus on LDS methods and the metrics used to evaluate them. Results indicate a rising adoption of hybrid models combining extractive and abstractive strategies, frequently powered by deep learning and optimization. Concurrently, evaluation practices have shifted from traditional overlap-based metrics (e.g., ROUGE) toward semantic measures such as BERTScore and MoverScore. However, these metrics still face challenges related to interpretability, domain adaptation, and computational cost. We advocate for the development of holistic, explainable, and reference-free evaluation frameworks aligned with human judgment to enhance the reliability and applicability of LDS systems across domains.

1. Introduction

With the exponential growth of scientific literature, there is an increasing demand for tools that can support the efficient synthesis of large volumes of textual information. Long Document Summarization (LDS) techniques have emerged as essential instruments for reducing time and effort in processes such as systematic literature reviews (SLR), where researchers must analyze and integrate knowledge from hundreds of documents. In domains such as medicine, law, and scientific publishing, LDS enables practitioners to extract relevant insights from lengthy and often highly technical documents. Consequently, the development of robust summarization methods and reliable evaluation metrics has become a critical area of research.

Summarization techniques are particularly useful when applied to large-scale databases like Scopus, Web of Science, or PubMed, where thousands of potentially relevant studies must be reviewed. By generating coherent and concise summaries, these methods facilitate rapid screening, topic modeling, and content synthesis, tasks that would otherwise require extensive human effort.

* Corresponding author at: Knowledge Discovery, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Schloss Birlinghoven 1, Sankt Augustin, 53757, Germany.
Email addresses: bady.gana@pucv.cl (B. Gana), hector.allende-cid@iais.fraunhofer.de (H. Allende-Cid), stefan.rueping@iais.fraunhofer.de (S. Rüping), marcelo.becerra@pucv.cl (M. Becerra-Rozas), juan.zamora@pucv.cl (J. Zamora).

https://doi.org/10.1016/j.neucom.2025.131287
Received 15 June 2025; Received in revised form 2 August 2025; Accepted 13 August 2025
Available online 22 August 2025

Several surveys have addressed various aspects of text summarization, particularly in the context of long documents and evaluation metrics. These surveys cover a wide range of approaches, from abstractive and extractive summarization to multi-document and query-focused summarization. For instance, Shakil et al. [1] provide an in-depth survey on abstractive summarization techniques, focusing on state-of-the-art methods and challenges, while also addressing long-document summarization and evaluation metrics. Koh et al. [2] focus on long document summarization, evaluating datasets, models, and metrics. Wahab et al. [3] offer a comprehensive review of optimization-based methods in extractive summarization, focusing on performance metrics. Additionally, Elsaid et al. [4] review Arabic text summarization techniques, emphasizing challenges such as dialects and morphological structure, and evaluating methods like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) for Arabic-specific summarization. Chaves et al. [5] analyze biomedical text summarization, highlighting specialized methods and metrics for this domain. Other surveys, such as those by Giarelis et al. [6] and Mishra et al. [7], review both extractive and abstractive summarization approaches and compare them using well-known metrics, while Alanzi and Alballaa [8] focus on query-focused multi-document summarization, offering insights into evaluation techniques for multi-document corpora.

The surveys by Meaney et al. [9] and Barbella et al. [10,11] investigate the evaluation of summarization models using metrics like ROUGE, with a particular focus on topic modeling and medical imaging summarization. Furthermore, Afsharizadeh et al. [12] and Aumiller et al. [13] provide critical reviews of multi-document summarization, with Aumiller's survey also evaluating German abstractive summarization models using ROUGE.

Hewapathirana et al. [14] provide an extensive review of multi-document summarization models, evaluate their performance on various datasets, and discuss future research directions. They use the ROUGE score for evaluation and contribute valuable insights for future MDS (Multi-Document Summarization) research. Similarly, Davoodijam et al. [15] categorize and analyze 42 metrics across dimensions such as intrinsic vs. extrinsic and manual vs. automatic, exploring challenges in their application.

By synthesizing these insights, this review not only maps the current landscape of multi-document summarization and evaluation metrics but also highlights key challenges and research gaps. Understanding how these methodologies and assessment frameworks evolve is crucial for advancing summarization techniques, particularly in evidence synthesis and large-scale document analysis. The following sections elaborate on the methodological approach taken in this study, the results obtained, and their broader implications for the field.

The remainder of this paper is organized as follows. In Section 2, we describe the methodology of this SLR, including the databases searched, inclusion and exclusion criteria, and data extraction procedures. Section 3 presents the main results, detailing the categorization of summarization methods and evaluation metrics. Section 4 offers a discussion of the findings, highlighting emerging trends, research gaps, and the limitations of current evaluation frameworks. Finally, Section 5 summarizes the key contributions of the study and outlines potential directions for future research in the field of long document summarization.

Contribution and innovation

This review differs from previous surveys by offering:

- A unified framework that connects LDS techniques with the evaluation metrics used to assess them.
- A focus on recent research trends (2022–2024), including deep learning-based models, transformer architectures tailored for long sequences (e.g., Longformer, BigBird), and optimization-based hybrid models.
- A comprehensive classification of evaluation metrics into lexical (e.g., ROUGE, BLEU), semantic (e.g., BERTScore, Moverscore), linguistic quality (e.g., readability, coherence), and human-based assessments.
- A forward-looking discussion on the limitations of current evaluation practices and the potential for *explainable*, *reference-free*, and *domain-adapted* metrics.

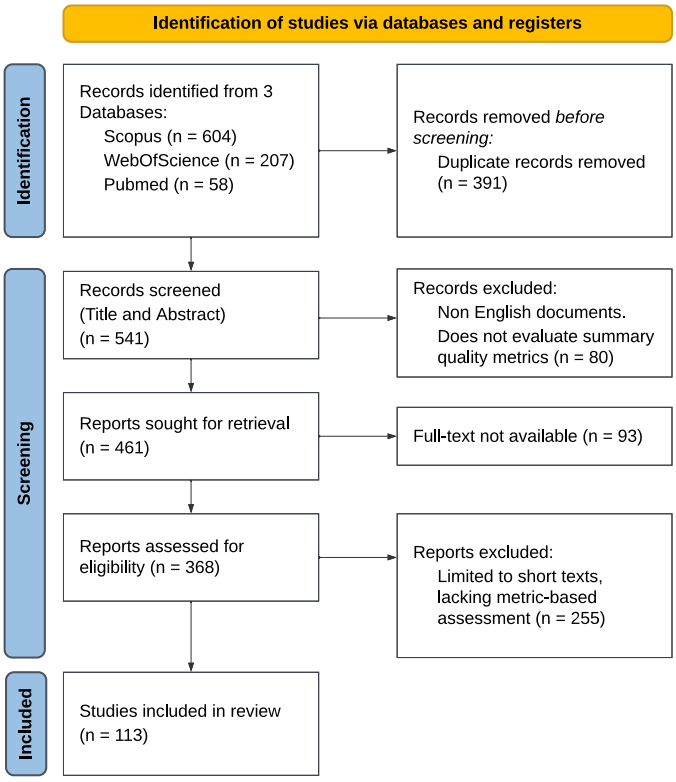


Fig. 1. PRISMA 2020 [16] Identification of studies via databases and registers.

2. Methodology

This SLR was conducted following the PRISMA 2020 guidelines, as shown in Fig. 1. The methodology was designed to ensure transparency, reproducibility, and rigor throughout the study identification, selection, and synthesis processes.

2.1. Research questions

To further understand the progress and current trends in this domain, the following research questions were explored:

1. **RQ1:** What are the main techniques used in long document summarization, and how are they categorized?
2. **RQ2:** What are the most commonly used evaluation metrics for assessing the quality of long document summaries in recent studies?
3. **RQ3:** What advancements in deep learning have been incorporated into long document summarization?
4. **RQ4:** Are there any limitations identified in recent literature regarding the metrics used to evaluate summary quality?

2.2. Eligibility criteria

The SLR was conducted using predefined inclusion and exclusion criteria to ensure the relevance and methodological robustness of selected studies.

Inclusion criteria:

- Articles published between January 1, 2022, and October 31, 2024.
- Studies focusing on long document or multi-document summarization and their evaluation.
- Peer-reviewed journal articles, conference papers, or proceedings written in English.
- Studies that explicitly discuss or apply evaluation metrics to assess summarization quality.

Exclusion criteria:

- Studies focusing exclusively on short or informal text summarization. We define short text summarization as approaches that target inputs that are very brief and structurally simple, such as tweets, user reviews, or isolated news headlines typically well under 750 tokens, and that often lack formal evaluation using established summarization metrics. These works are excluded due to their limited relevance to the methodological and evaluative scope of this study.
- Articles lacking explicit evaluation metrics or detailed methodological descriptions.
- Studies for which full text was not available.

2.3. Information sources

Three academic databases were used to identify relevant literature:

- **SCOPUS:** 604 initial results.
- **Web of Science (WOS):** 270 initial results.
- **PubMed:** 58 initial results.

Backward reference searching was also conducted to identify additional relevant studies. The final search was completed on October 31, 2024.

2.4. Search strategy

To ensure both the breadth and precision of the retrieved literature, the search strategy combined Boolean operators with domain-specific terminology. These resulted from a preliminary mapping of key concepts identified in recent high-impact publications on long document summarization. The search focused on three main dimensions: the nature of the input, such as long or multi-source documents; the methodological approach, including automatic, supervised and unsupervised techniques; and the evaluation strategies most commonly reported. The terminology was iteratively refined to reflect the prevailing discourse in the field and to align with the review's analytical objectives. Although the query explicitly included evaluation-related terms and an English-language filter, database indexing inconsistencies produced records that either did not truly address evaluation practices or were not in English. These cases were resolved during screening using the conservative rule described in Section 2.5. The exact search string used was:

```
("document" OR "documents" OR "long documents"  OR "large documents"
OR "scientific documents" OR "medical documents" OR "multi-document"
OR "multiple documents") AND (``summarization" OR "automatic summarization"
OR "supervised summarization" OR "unsupervised summarization" OR "text
summarization" OR "document summarization" OR "corpus summarization")
AND ("evaluation metrics" OR "summarization evaluation" OR "performance
metrics" OR "text quality metrics" OR ``ROUGE" OR "BLEU" OR ``BERTScore"
OR ``coherence" OR ``fidelity" OR ``relevance" OR "unsupervised evaluation"
OR "supervised evaluation")
```

Filters were applied to restrict results to English-language publications and the specified publication window.

2.5. Selection process

Following PRISMA 2020 and the inclusion/exclusion criteria defined in Section 2.2, we processed 932 records (Scopus = 604, Web of Science = 270, PubMed = 58). After removing 391 duplicates and non-DOI entries, 541 unique records remained.

Two independent reviewers screened titles and abstracts. A record was excluded at this stage only when both reviewers agreed that it was not in English and that neither the title nor the abstract contained any explicit or implicit indication of evaluation practices (e.g., "metrics", "assessment", "performance evaluation", "quality criteria"). Ambiguous cases were conservatively advanced to full-text review to minimize false negatives and selection bias. This step resulted in 80 exclusions and 461 full-text retrieval attempts, of which 93 were unavailable due to access issues.

From the 368 full texts assessed, 255 were excluded because they focused on short-text summarization or lacked substantive treatment of evaluation metrics, leaving 113 studies for synthesis.

We piloted the screening criteria on a random subset to calibrate decisions. Disagreements were resolved by consensus, always favoring inclusion in doubtful cases. Although we did not compute an inter-rater statistic (e.g., Cohen's kappa), the dual-review plus consensus procedure balanced feasibility and bias control.

2.6. Data collection process

We used a structured extraction form to ensure consistency. The following fields were captured:

- **Bibliographic data:** title, authors, year, publication type.
- **Study context:** summarization type (single/multi/hybrid), application domain, and source dataset.
- **Evaluation methodology:** automatic metrics (e.g., ROUGE, BLEU, BERTScore), interpretative dimensions (coverage, redundancy, coherence), rationale, and reported limitations.

All information was consolidated in a master spreadsheet to support qualitative coding and frequency-based analyses. Deduplication by DOI and data cleaning were automated with Python/Pandas scripts to guarantee traceability and exclude ambiguous or non–peer-reviewed records. Quality checks (e.g., controlled vocabularies and spot audits) were applied to reduce extraction errors.

2.7. Synthesis methods

Due to the heterogeneity in study types and reported outcomes, a narrative synthesis approach was adopted. This included:

- Grouping studies by summarization type (extractive, abstractive, hybrid) and evaluation metrics.
- Mapping metric usage frequency and domain-specific applications.
- Highlighting challenges, such as reproducibility issues and over-reliance on surface-level metrics.

Visualizations were used to illustrate metric distributions and emerging trends.

2.8. Certainty assessment

A qualitative certainty assessment was conducted, emphasizing:

- Methodological transparency and reproducibility of the studies.
- Consistency in findings across similar domains.
- Identified biases or gaps, such as domain underrepresentation and limited diversification of evaluation metrics beyond lexical-overlap measures.

This assessment informed the interpretation of evidence strength and helped identify research gaps for future exploration.

3. Results

This section presents the main findings of our SLR on Long-Document Summarization (LDS). Through an extensive analysis of recent works, we structured the field into a set of interconnected components that reflect the current research landscape. These components include the types of

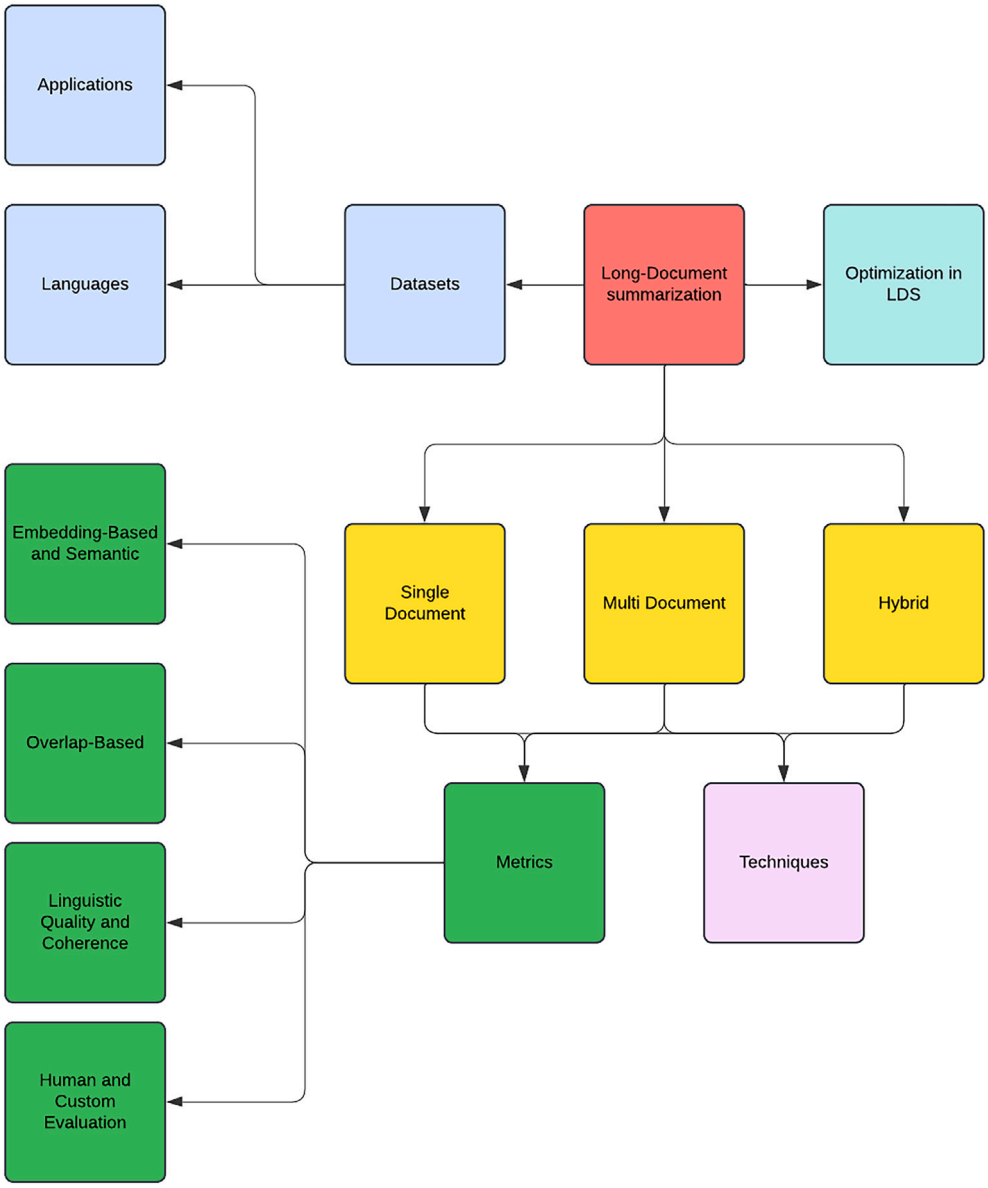


Fig. 2. Structure of the systematic literature review on Long-Document Summarization, including datasets and their associated languages and applications, identified summarization approaches (single-document, multi-document, and hybrid), and their corresponding techniques and evaluation metrics. Optimization strategies in LDS are also included.

datasets employed, the languages and application domains addressed, the summarization methods applied, and the techniques and evaluation metrics used. Furthermore, optimization strategies aimed at improving LDS outcomes are also highlighted.

Fig. 2 provides a visual summary of the structure derived from our SLR on LDS. The diagram highlights key dimensions explored in the field, including the datasets used, along with the languages and applications they support.

We identify three primary types of summarization methods: single-document summarization, multi-document summarization, and hybrid approaches that incorporate elements of both. Fig. 3 presents the distribution of methods across these summarization types. Most research focuses on single-document summarization, with a substantial portion employing purely extractive techniques. Multi-document summarization follows in volume, also predominantly extractive, while hybrid approaches (i.e., applicable to both single and multi-document settings) remain comparatively less explored. Across all categories, purely extractive methods are more prevalent, though there is a noticeable presence of abstractive and mixed-method approaches in single-document tasks.

Each method is associated with specific techniques and evaluation metrics. These metrics are categorized into four major groups: embedding-based and semantic, overlap-based, linguistic quality and coherence, and human/custom evaluation.

Additionally, the diagram includes a branch dedicated to optimization strategies applied in LDS. This holistic overview provides insights into the current landscape and ongoing research directions in LDS.

3.1. Datasets

The datasets analyzed in this systematic literature review encompass a wide range of text summarization applications, spanning various domains, languages, and years of publication in Table 1. These datasets serve as benchmarks for evaluating summarization models, allowing researchers to test and compare methodologies across different contexts. Below, we provide a chronological overview of

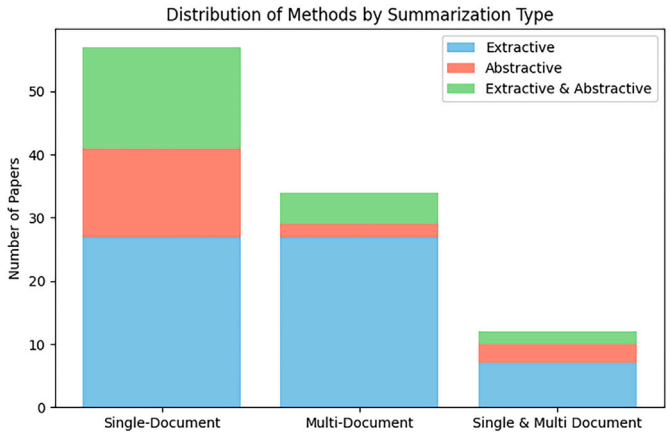


Fig. 3. Distribution of Methods by Summarization Type.

the evolution of these datasets and their impact on summarization research.

The development of summarization datasets began in the early 2000s with the Document Understanding Conference (DUC) series, which established a foundational framework for news and general document summarization. Datasets such as DUC 2002, DUC 2003, and DUC 2004 [17–22] introduced multi-document summarization tasks, enabling systematic evaluation of summarization models. These were followed by DUC 2005, DUC 2006, and DUC 2007 [23–29], which further refined the methodologies by incorporating more extensive documents and well-defined summarization objectives. In 2008 and 2009, the Text Analysis Conference (TAC) expanded upon these efforts by introducing new evaluation tracks aimed at identifying key information across related document sets [30–32].

By 2015, the field witnessed a paradigm shift with the introduction of large-scale datasets, such as CNN/DAILYMAIL [22,33–36], which facilitated the transition to neural summarization models by providing millions of article-summary pairs. Simultaneously, the emergence of CL-SCISUMM 2016 [37] marked the beginning of scientific summarization research, focusing on extracting relevant information from academic literature.

Between 2018 and 2021, summarization datasets diversified, reflecting the increasing specialization of tasks. In the biomedical domain, datasets such as PUBMED [38–40] and BioMedAC [17] became instrumental in summarizing medical literature, enhancing accessibility to critical diagnostic and treatment-related information. Legal summarization also gained traction with datasets such as BILLSUM [41–44], CAIL-2020 [45–47], and GOVREPORT [41,48,49], which enabled the development of models capable of synthesizing complex legal and governmental texts. Additionally, the introduction of MULTI-NEWS [22,36,50,51] advanced multi-document summarization, challenging models to aggregate and distill information from multiple sources.

During this period, new summarization challenges emerged across various text genres. WIKISUM [22,70,78] and WikiHow [46,83] explored summarization in encyclopedic and instructional content, requiring models to adapt to different writing styles. Meanwhile, Debatepedia [80] and Opinosis [46] facilitated the summarization of user-generated content, debates, and opinions, while financial datasets such as FINDSum [84] and ECTSum [93] focused on extracting key insights from corporate reports and earnings call transcripts. The field also witnessed notable advancements in multilingual summarization, with datasets such as Mawdoo3 (Arabic) [44], Hindi Literature [28], and CAIL-2020 (Chinese legal documents) [88], which highlight the complexity of adapting summarization techniques to varied linguistic structures and domain-specific conventions. These resources not only broaden the geographical and linguistic scope of LDS research, but also bring to light fundamental challenges in achieving robust cross-lingual performance, challenges that recent datasets such as LoRaLay, CLSum, and MLSum further exemplify through empirical evidence. Building on this trend, recent datasets such as LoRaLay, CLSum, and MLSum demonstrate that linguistic diversity impacts summarization models beyond syntactic variation [109], including differences in layout, rhetorical structure, and domain-specific terminology. Specifically, Shakil et al. [1] emphasize that cross-lingual summarization involves risks of semantic loss and structural mismatches during translation. While common pipeline approaches that translate first then summarize may propagate errors, integrated neural machine translation and summarization models better preserve semantic integrity. CLSum reveals poor generalization across legal systems, and LoRaLay highlights performance drops on mixed-language documents, underscoring the necessity for multilingual encoders and alignment mechanisms sensitive to structural and semantic heterogeneity.

These challenges are compounded in low-resource languages and specialized domains like legal summarization, where annotated data is scarce and costly. Techniques such as transfer learning, back-translation, and knowledge-guided data augmentation help mitigate these limitations. For instance, Nguyen et al. [109] introduce a rephrasing method using legal constraints in LLM prompting to generate

Table 1
Summary of most common datasets used in text summarization tasks, including language, year, application domain, and relevant papers.

Dataset	Language	Year	Application	Papers
DUC 2002	English	2002	General, News	[17,18,20,21,50,52–56]
DUC 2003	English	2003	General, News	[18–20,53,57]
DUC 2004	English	2004	General, News	[18–22,33,34,50,53,56,58–61]
DUC 2005	English	2005	General, News	[20,23–25,57,62,63]
DUC 2006	English	2006	General, News	[20,23,24,26,27,62–64]
DUC 2007	English	2007	General, News	[20,23,24,26,28,29,62]
CNN/DAILYMAIL	English	2015	News, Media	[19,21,22,33–35,50,51,58,65–72]
PUBMED	English	2018	Biomedical, Medical Research	[38–40,48,73]. [49,67,74,75]
ARXIV	English	2019	Scientific Articles	[39,48,49,67,74,76,77]
CL-SCISUMM 2016	English	2016	Scientific Analysis	[37]
BILLSUM	English	2019	Legal, Government Documents	[41–44]
TAC2008	English	2008	General, News	[31]
TAC2009	English	2009	General, News	[30–32]
CAIL-2020	Chinese	2020	Judicial Summarization, Legal	[45–47]
WIKISUM	English	2018	General Topics, Encyclopedic	[22,70,78]
GOVREPORT	English	2020	Government Reports	[41,48],[49]
MULTI-NEWS	English	2019	Multi-document, News	[22,36,50,51]
ML-SUM	English	2021	Machine Learning Research	[44,46]
Other	Various	Various	Miscellaneous	[17,19,28,37,38,42,44–46,60,62,65,66,70,76,79–113]

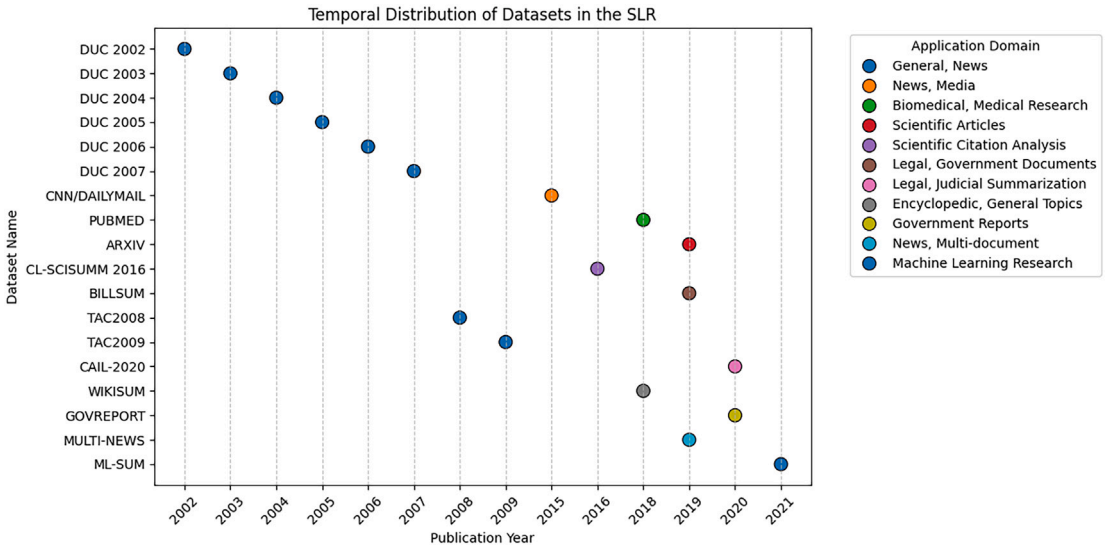


Fig. 4. Temporal Distribution of common Datasets.

synthetic data with improved domain fidelity, which is particularly beneficial in data-scarce settings. Evaluation remains a critical bottleneck. Standard metrics like ROUGE often fail to capture semantic adequacy and domain relevance in multilingual outputs. To address this, LTScore [112] incorporates legal knowledge via fine-tuned models, emphasizing key terms and valid paraphrases. The lack of benchmarks in languages such as Urdu further reinforces the urgent need for multilingual, semantically-aware evaluation frameworks.

These datasets not only address language-specific nuances but also contribute to domain adaptation, ensuring the applicability of summarization models in specialized fields. Additionally, several datasets cater to niche applications, including DailyMed for drug label summarization [101], EDUVSUM for educational video summaries [94], and ClueWeb09 for webpage summarization [100]. Large-scale newswire datasets such as Robust04 [87] and GOV2 [87,100] further support research in text retrieval and automatic summarization, broadening the scope of real-world applications.

Fig. 4 illustrates the evolution of the field, from early general-purpose summarization benchmarks (e.g., DUC and TAC series) to more recent datasets focused on specialized domains such as biomedical research (PUBMED), scientific literature (ARXIV), and legal documents (BILLSUM, CAIL-2020). This temporal progression highlights both methodological advancements and the growing diversification of application domains within summarization research.

A significant gap is observed between 2009 and 2015, reflecting a period with limited publicly released datasets. This may be attributed to a transitional phase in the field, marked by shifts from extractive to abstractive methods and the growing reliance on proprietary or private data sources during the emergence of neural approaches. The reactivation of dataset publication around 2015, starting with CNN/DAILYMAIL, coincides with the rise of neural sequence-to-sequence models and large-scale pretrained language models.

3.2. Extractive and abstractive summarization techniques for single-document texts

Automatic single-document summarization has been extensively studied, with approaches primarily categorized as extractive and abstractive. Extractive methods select key sentences directly from the original text to compose a summary, whereas abstractive methods generate new sentences to express the information in a more concise and natural manner.

Extractive summarization methods aim to identify and retain key sentences from the original document while maintaining their structure. GETS (Graph-based Extractive Text Summarization Sentence Scoring Scheme) [79] employs graph-based sentence selection, where sentences are treated as nodes, and Jaccard similarity is used to establish relationships between them. The method enhances coherence through graph clustering, ensuring that the extracted sentences maintain contextual relevance. In legal text summarization, Bayesian Optimization [41] integrates Latent Dirichlet Allocation (LDA) and contextual sentence embeddings from LegalBERT to refine the selection of representative sentences. In the same domain, DCESumm [43], integrates supervised sentence prediction with deep clustering techniques to enhance extractive summarization. Other extractive techniques, such as SeburSum [65], leverage contrastive learning with BERT and RoBERTa, reframing extractive summarization as a semantic text matching problem. Meanwhile, BERTSUM models [83] apply document partitioning and clustering to handle lengthy legal documents effectively, ensuring that key points are retained without loss of coherence.

In contrast, abstractive summarization reconstructs a document's main ideas using newly generated text, requiring a deeper understanding of semantic relationships and linguistic structures. EFAS (Entity-Driven Fact-Aware Summarization) [38] integrates UMLS, ICD-10, and SNOMED-CT knowledge bases with BioBERT embeddings, enhancing biomedical text summarization while maintaining factual integrity. Extract-then-Assign (ETA) [42] improves abstractive summarization in legal texts by first generating an extractive summary, which is then refined using BART fine-tuning. In financial summarization, FINDSum [84] employs a structured approach that combines Generate-then-Combine (GC) and Generate-Template-then-Fill (GTF) techniques to ensure numerical accuracy in generated summaries.

Hybrid summarization methods integrate extractive and abstractive elements to improve coherence and readability. CNN-GRU-based hybrid models [58] utilize reinforcement learning to optimize the extraction and generation processes while leveraging Word2Vec and BERT embeddings for improved word representations. In legal text summarization, Lawformer [88] combines BERTSUM-based extraction with a Pointer-Generator Network, ensuring that summaries are both legally sound and contextually accurate.

Graph-based summarization techniques have also gained traction, particularly for improving structure and coherence. Multi-granularity heterogeneous graph models [66] establish hierarchical relationships between words, sentences, and topics, enhancing sentence selection.

Table 2
Classification of summarization techniques.

Technique	References
Graph-based and topic-aware summarization	[40,49,66,74,76,79]
Extractive and hybrid summarization	[39,41,54,58,65,69,70,88]
Abstractive summarization and deep learning	[38,43,71–73,87,95,116]
Reinforcement learning and optimization	[48,55,93,97]
Domain-specific summarization (legal, biomedical, financial)	[84,89,98,113]

Similarly, GoSum [74] integrates reinforcement learning with Graph Neural Networks to optimize scientific text summarization. Additionally, graph-based abstractive summarization models (GBAS) [76] employ SciBERT and Graph Transformer Networks (GTN) to construct structured knowledge representations, improving summary accuracy and contextual understanding.

Deep learning techniques continue to drive innovation in summarization. SE-BERT [87] extends BERT's capabilities by incorporating abstractive text generation with PEGASUS, while ODL-LTS (Optimal Deep Learning-Based Legal Text Summarization) [95] applies TF-IDF, Rouge-L similarity measures, and a glowworm swarm-optimized BiGRNN model to refine legal text summarization. Additionally, multi-head self-attention mechanisms have been incorporated into pointer network-based summarization [71] to improve fluency and coherence. Other hybrid deep learning models, such as the two-phase deep neural document summarization framework [72], combine extractive intra-cosine attention similarity with BiLSTM-based abstractive summarization, demonstrating the effectiveness of hybrid models.

Optimization and reinforcement learning strategies further refine summarization methodologies. For example, UOTSumm [48] leverages Unbalanced Optimal Transport (UOT) to align document sections and summary sentences optimally. FLAN-FinBPS [93] integrates an unsupervised question-based context generator with a supervised instruction-tuned summarization model to improve financial document summarization. In the energy sector, aspect-based extractive summarization models [97] utilize MapReduce and BERT, akin to the large-scale approach of Leiva-Araos et al. [114] with K-means clustering to enhance domain-specific summarization. Some models even incorporate game theory principles, where sentence selection is modeled as a strategic interaction between sentences, using replicator dynamics to optimize representation [55].

Domain-specific summarization methods continue to evolve, adapting to specialized needs. LegalSumm [98] generates multi-perspective legal case summaries, ensuring that different viewpoints are preserved. In biomedical contexts, MedicoVerse [73] employs SapBERT embeddings, hierarchical clustering such as [115], and abstractive summarization to enhance medical text summarization. Similarly, hospital discharge summaries [89] incorporate metadata such as disease type, physician details, and patient length of stay, using Longformer-based architectures. In legal text processing, models based on Cross Latent Semantic Analysis (CLSA) with LSTM [113] extract key sentences from court decisions, demonstrating deep learning adaptability in judicial contexts.

These diverse methodologies illustrate the breadth of single-document summarization techniques, which span extractive, abstractive, hybrid, and graph-based models. The growing integration of deep learning, reinforcement learning, and optimization continues to push the field forward, enhancing adaptability and robustness across various domains. A structured overview of these methodologies, categorized based on their core techniques, is provided in Table 2, highlighting key references for each summarization approach.

3.2.1. Metrics

Evaluating single-document summarization requires assessing multiple factors such as informativeness, coherence, and fluency. To achieve this, different types of metrics have been employed, broadly categorized into overlap-based metrics, embedding-based metrics, linguistic quality metrics, and human evaluation methods.

Overlap-based. The most commonly used overlap-based metric is ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), which quantifies unigram, bigram, and longest common subsequence overlaps [19,68,70]. These methods provide a straightforward way to assess content similarity and have been extensively applied to both extractive and abstractive summarization models. A more flexible variant, ROUGE-SU4, evaluates skip-bigram co-occurrences to capture looser word orderings and has also been used in several studies [59,63]. For a detailed formal description of these metrics, see Section 4.2.

Embedding-based and semantic. While overlap-based metrics provide a basic assessment of content retention, they do not capture semantic meaning. To address this limitation, embedding-based metrics evaluate how similar summaries are at a deeper linguistic level. BERTScore [38,42,48] measures similarity by comparing contextual word embeddings rather than exact word matches. MoverScore [41] expands on this by computing the semantic distance between words using pre-trained word embeddings, ensuring better alignment of meaning. Other common semantic evaluation methods include TF-IDF similarity and cosine similarity [19,95], which help assess sentence relevance. Additionally, LDA [41] has been applied within Bayesian Optimization frameworks to evaluate topic consistency in summarization.

Linguistic quality and coherence. Beyond textual similarity, assessing the linguistic quality of a summary is essential for readability and coherence. Readability indices such as Flesch-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS) [110] measure the ease of understanding a text. In addition, Maximum Marginal Relevance (MMR) [37,100] is used to balance summary diversity and relevance, reducing redundancy while ensuring comprehensiveness. Some approaches employ graph-based embeddings like Node2Vec and DeepWalk to assess text coherence, capturing relationships between sentences and paragraphs [91].

Human and custom evaluation. Despite advancements in automated evaluation, human assessment remains a gold standard for evaluating summarization quality. Human evaluation typically includes expert ranking and manual assessment of summaries based on informativeness, fluency, conciseness, and coherence [52,53,59,63,78,80,81]. Likert-scale ratings and comparisons with gold-standard summaries offer quantitative assessments [31,32,34,44,50,52,53,59,63,64,81,85,99,106]. Some studies integrate domain-specific evaluations, such as physician-based assessments for medical summaries [89,116], legal expert evaluations for legal text summarization [98], and business-oriented KPIs for summarization quality [73]. Additionally, question-answering-based assessments have emerged as a novel method to determine how well summaries retain key information [78].

Fig. 5 presents the frequency of metric usage across three types of single-document summarization approaches: extractive, abstractive, and mixed. As shown, overlap-based metrics such as ROUGE-1, ROUGE-2, and ROUGE-L are by far the most commonly used across all summarization types, especially in single extractive summaries, where each is employed in over 19 studies. In contrast, more semantically oriented metrics such as BERTScore are rarely used, appearing in only 4 abstractive and 3 mixed summarization studies. Similarly, precision, recall, and F-score are sporadically reported, and linguistic quality metrics such as ROUGE-SU4 and ROUGE-3 are scarcely mentioned.

These findings suggest a strong reliance on traditional lexical overlap metrics, particularly in extractive approaches, while more advanced or

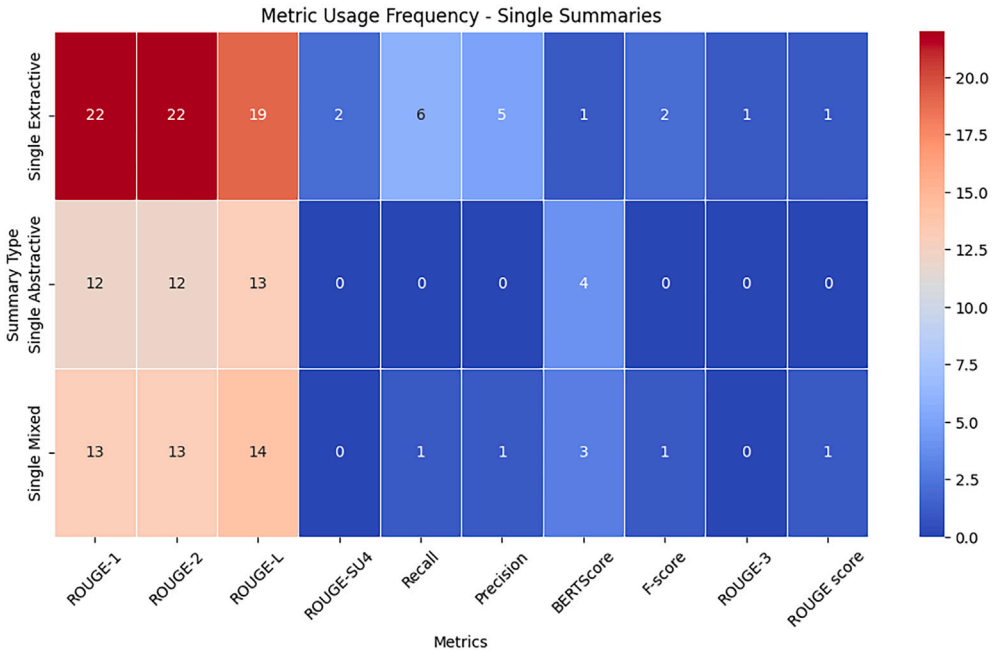


Fig. 5. Most frequent metrics for single documents summarization.

semantically nuanced metrics remain underutilized in single-document contexts. This pattern highlights a potential gap in evaluation practices, especially given that ROUGE metrics may not fully capture the semantic adequacy and factual consistency of abstractive summaries. The results reinforce the need for more diverse and semantically aware evaluation frameworks tailored to the specific demands of single-document summarization.

3.3. Extractive and abstractive summarization techniques for multi-document texts

Summarizing multiple documents introduces complexities beyond those encountered in single-document summarization. The main challenge lies in integrating information from diverse sources while ensuring coherence, minimizing redundancy, and maintaining comprehensive topic coverage. Various techniques have been proposed to address these challenges, leveraging different strategies for selecting and generating content.

One effective approach to structuring extractive multi-document summaries involves rhetorical and graph-based methods. Grapharizer [59] integrates word graphs and topic modeling using LDA to capture relationships across documents. SGCSumm [33] employs graph convolutional networks (GCN) to build context-aware representations of document content, refining sentence selection through a structured ranking process. In addition to graph-based techniques, machine learning-driven extractive models, such as evolutionary sparse multi-objective algorithms [52] and Firefly-based selection strategies [53], optimize sentence inclusion by balancing informativeness and diversity. CohQFMDS-Sum [63] offers an unsupervised framework capable of summarizing text from varied sources without requiring predefined labeled datasets. This adaptability makes it particularly effective in handling heterogeneous data, where training samples may not always be available. Similarly, clustering techniques, including hierarchical tree-based models [18] and language-independent clustering approaches [44], further refine summary selection by grouping similar content and minimizing redundancy.

In the abstractive context EDITSum [81] enhances this process through hierarchical decoding and VAE, capturing nuanced relationships between sentences. In cases where discourse plays a central role, models like TOMDS [78] integrate topic-aware attention mechanisms and discourse parsing, restructuring text to improve readability. Similarly, RSGen [80] enhances coherence by incorporating rhetorical structures into both encoding and decoding processes. It employs a graph transformer to encode relationships between text elements, followed by a structured rhetorical plan in the decoding stage to ensure logically ordered summaries. Hybrid frameworks merge the strengths of both extractive and abstractive paradigms. HMSumm [50] extracts salient content using deep submodular networks (DSN) and refines it with BART and T5 models, while PDSum [106] employs contrastive learning to iteratively enhance document representations, adapting summaries dynamically to evolving datasets.

Summarization can also be approached as an optimization challenge, where methods aim to maximize informativeness while controlling redundancy. Multi-objective optimization strategies, such as Multi-Objective Shuffled Frog-Leaping Algorithm (MOSFLA) [30] and Multi-Objective Number-One-Selection Genetic Algorithm (MONOGA) [31], use evolutionary computing to refine sentence selection iteratively. In more targeted applications, Indicator-based Multi-Objective Variable Neighborhood Search (IMOVNS) [32] incorporates mutation and repair techniques to optimize summaries for query-focused retrieval. Similarly, Multi-Objective Ant Colony Optimization (MOACO) [24] applies Pareto Ant Colony Optimization (P-ACO) to explore diverse sentence combinations, prioritizing content based on multiple evaluation criteria.

Beyond traditional methodologies, emerging summarization techniques introduce novel perspectives. Semantic-driven approaches, such as SDbQfSum [64], enhance topic-based summaries by leveraging Wikipedia commonsense knowledge, ensuring that generated content aligns with contextual meaning. Discourse-aware models [36] construct discourse trees to establish logical relationships between textual elements, improving summary cohesion. Readability-focused methods, including K-means clustering with Flesch readability scoring [25], optimize summaries for accessibility, making them more digestible for broader audiences. Additionally, language-specific models like AraTSum [92] adapt summarization strategies to Arabic-language Twitter discussions, demonstrating the adaptability of these techniques across linguistic domains.

Table 3
Classification of multi-document summarization techniques.

Technique	References
Graph-based and structural summarization	[33,59,63,78,80]
Extractive and hybrid summarization	[18,33,44,50,52,53,59,106]
Abstractive summarization and deep learning	[36,64,78,80,81]
Optimization-based and evolutionary algorithms	[20,24,30–32]
Domain-specific and query-focused summarization	[25,26,92,107]

Other summarization strategies refine output quality through ranking and user interaction. Maximal gSpan [60] applies directed co-occurrence graphs to prioritize frequently referenced terms, facilitating structured summary generation. Position-based ranking models, such as those introduced in [26], weight sentences based on their placement within documents, ensuring that key information is highlighted. Interactive query-based summarization, as demonstrated by genetic algorithm-driven frameworks [107], incorporates user feedback to dynamically refine summaries, allowing for adaptive, user-guided results.

As research progresses, the integration of structured extraction, abstractive synthesis, and optimization techniques continues to expand the scope of multi-document summarization (see Table 3). These advancements reflect an ongoing effort to develop methods that are not only accurate and coherent but also adaptable to diverse domains and evolving information needs.

3.3.1. Metrics

Evaluating multi-document summarization involves several key aspects, including informativeness, coherence, and redundancy reduction. Metrics are categorized into four main types: overlap-based metrics, embedding-based metrics, linguistic quality metrics, and human evaluation methods.

Overlap-based. These metrics assess how much textual content from a generated summary overlaps with reference summaries. The most widely used measures include ROUGE-1, ROUGE-2, and ROUGE-L, respectively [32,76]. Some studies extend evaluation with ROUGE-SU4, which accounts for skip-bigram matches to capture non-contiguous word relationships [59,63].

Embedding-based and semantic. To go beyond direct textual overlaps, embedding-based metrics evaluate semantic similarities between system-generated and reference summaries. BERTScore and Sentence-BERT [61,63,106] analyze contextual embeddings to measure similarity at a deeper linguistic level. Other semantic evaluation methods include TF-IDF similarity, explicit semantic analysis (ESA), and Jaccard Index, which help detect redundant or unrelated content [18,30,34,44,64,85].

Linguistic quality and coherence. Linguistic quality metrics focus on the readability and coherence of summaries. Readability indices such as FKGL and Gunning Fog Score [25] measure how easy the text is to understand. Some studies use neural coherence models trained with contrastive learning [80] to evaluate how well sentences flow logically. Sentence reordering algorithms, such as chronological ordering and topic closeness ranking, further refine the organization of multi-document summaries [26,63].

Human and custom evaluation. Despite advancements in automated evaluation, human assessment remains a crucial benchmark for summarization quality. Expert annotators rank summaries based on informativeness, fluency, and coherence [52,53,59,63,78,80,81]. Some studies utilize Likert-scale ratings and compare system-generated summaries with gold-standard references [31,32,34,44,50,52,53,59,63,64,81,85,99,106]. Additionally, domain-specific evaluations, such as medical expert reviews for clinical document summarization [89,116] and legal professional assessments for legal texts [98], provide more tailored insights into summarization effectiveness. Another emerging evaluation method involves question-answering-based assessments, which measure how well key information is preserved in the generated summary [78].

Fig. 6 shows the frequency of metric usage across multi-document summarization approaches: extractive, abstractive, and mixed. Similar to single-document summarization, ROUGE-1, ROUGE-2, and ROUGE-L remain the most frequently used metrics, especially in extractive methods, where ROUGE-1 reaches a frequency of 22. However, compared to single-document summarization, the usage of alternative metrics such as

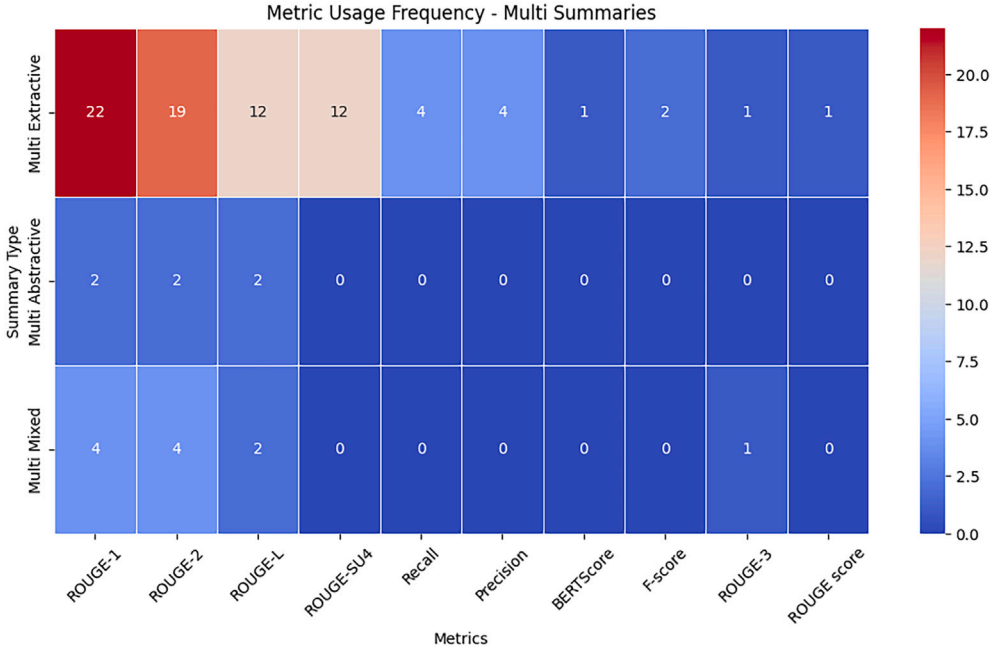


Fig. 6. Most frecuent metrics for multi documents summarization.

BERTScore, Recall, and Precision is even more limited. Multi abstractive and mixed approaches display very low metric diversity and frequency. Notably, no semantic or embedding-based metrics were reported in these types, and most of the metric use is concentrated in extractive methods. ROUGE-SU4 appears with moderate frequency (12 times) only in extractive summaries, reinforcing the dominance of overlap-based evaluation strategies. Compared to the single-summary scenario (Fig. 5), where there is a slight increase in metric variety in mixed or abstractive approaches, the evaluation of multi-document summaries remains highly conservative. This suggests a lag in adopting more sophisticated or semantically-aware metrics in the multi-document context, despite the greater complexity and redundancy challenges posed by these summarization tasks. Overall, the results highlight the need for improved evaluation practices that can better reflect the unique challenges of multi-document summarization, particularly for abstractive and hybrid models.

Overlap-based metrics, such as ROUGE, are predominantly utilized, often overshadowing semantic-based measures in both scenarios, i.e., single and multi-document. This prevalence can be attributed to their conceptual simplicity, computational efficiency, and historical precedence in the field. Their low computational cost, for instance, permits researchers to perform numerous evaluations during model tuning without significant resource constraints. Furthermore, and perhaps more critically, their historical precedence creates a strong inertia; new summarization methods must be compared against a vast body of literature that uses these established metrics, making the adoption of novel evaluation techniques challenging.

3.4. Hybrid summarization approaches and new metrics for sentence selection

The field of automatic text summarization has witnessed significant advancements by integrating single-document and multi-document summarization techniques into hybrid models. These approaches aim to leverage the strengths of both paradigms, ensuring that generated summaries are coherent, contextually relevant, and capable of handling diverse sources of information. These summarization techniques employ deep learning models, optimization algorithms, and linguistic analysis.

Pre-trained transformer models have played a crucial role in hybrid summarization. The model proposed in [62] utilizes BERTSUM, where BERT serves as the encoder, and a Transformer-based decoder generates abstractive summaries. The model employs domain adaptation techniques, including transfer learning and weakly supervised learning, to enhance summarization performance across different datasets, effectively bridging the gap between single-document and multi-document summarization.

Document Vector Method: An alternative approach, introduced in [28], employs document embeddings to capture semantic relationships among sentences. By clustering similar sentences using K-means, this model enhances the relevance and coherence of summaries. Sentence ranking is determined based on redundancy rate, diversity, and compression rate, ensuring that generated summaries retain key information without excessive repetition.

Unsupervised Extractive summarization model in [82] integrates concept maps with the RAKE (Rapid Automatic Keyword Extraction) method for single-document summarization, while applying Latent Semantic Analysis (LSA) for multi-document summarization. This unsupervised technique emphasizes keyword extraction and semantic similarity to enhance summary informativeness and coherence.

Advancements in large language models (LLMs) have enabled more powerful summarization techniques. The system in [90] employs the Mistral7B model for privacy policy summarization, utilizing BASE analysis for single-document summaries and WRT and REV techniques for cross-document insights. By incorporating prompt chaining, the model ensures accurate extraction of essential content while preserving readability and explainability.

Topic Modeling with Latent Dirichlet Allocation The framework in [75] integrates topic modeling with LDA to extract key themes from medical research articles. By scoring sentences based on relevance, this model identifies commonalities across documents and produces structured summaries that preserve domain-specific terminology.

Heterogeneous Graph Neural Networks Graph-based approaches continue to gain traction in hybrid summarization. The multi-granularity adaptive summarization model in [51] employs a heterogeneous graph neural network, incorporating TF-IDF, LDA topic modeling, and GloVe embeddings to establish meaningful relationships between sentences and topics, thereby improving summary coherence.

SMATS Model with Optimization Algorithms has been widely adopted for hybrid summarization. The SMATS model in [57] utilizes extractive summarization in combination with optimization algorithms such as Ant Colony Optimization (ACO), Bat Algorithm (BA), Cuckoo Search Optimization (CSO), Firefly Algorithm (FA), and Flower Pollination Algorithm (FPA). This approach optimizes sentence selection, producing concise and informative summaries.

3.4.1. Metrics

Evaluating hybrid summarization models requires a combination of automated and human assessment methods, to ensure that the summaries are accurate and coherent across different domains.

Overlap-based. Overlap-based metrics such as ROUGE remain a standard evaluation method for text summarization. Research in [28] reports Precision, Recall, and F-score, highlighting the effectiveness of ROUGE across different datasets, including MS-MARCO and DUC [62,82]. Additionally, ROUGE-3 and ROUGE-S are explored in [75] to provide more granular evaluations.

Embedding-based. To assess semantic similarity beyond n-gram overlaps, models employ embedding-based metrics such as RoBERTa fine-tuned for sentence similarity [62] and Doc2Vec [28] for capturing semantic relations. Cosine similarity between sentence embeddings is widely used [62,82], while LSA helps evaluate semantic consistency [75,82]. More advanced techniques include topic-based graph embeddings [51] and deep learning-based metrics, such as those implemented in the Mistral7B model [90].

Linguistic quality. Assessing linguistic quality and readability is crucial for summarization effectiveness. Studies evaluate cohesion and fluency using readability metrics such as FKGL, Gunning Fog Score, and SMOG Index [57]. Redundancy removal techniques, such as n-gram overlap filtering [28,82], help ensure conciseness, while concept maps [82] enhance coherence. Prompt chaining, as used in [90], further refines summarization by improving decision-making and explainability.

Human evaluation. Human evaluation remains indispensable for validating the quality of hybrid summarization models. Studies employ expert reviewers and domain specialists to assess summaries based on coherence, fluency, and informativeness. In [62], human judges provide scores on a 1-to-5 rating scale, while [82] collects evaluations from domain-specific experts. Other studies, such as [90], perform manual assessments of privacy policy summaries, comparing system-generated summaries with human-annotated references. Additionally, datasets like DUC-2003 and DUC-2005 [51,57] are used for benchmarking hybrid summarization models against expert-curated summaries.

These diverse methodologies illustrate the continuous advancements in hybrid summarization, integrating extractive, abstractive, and optimization-based approaches to improve text summarization across multiple domains. The combination of structured evaluation techniques ensures that hybrid models produce summaries that are both informative and coherent, bridging the gap between single-document and multi-document summarization.

3.5. Optimization approaches and user-based learning

As summarization techniques continue to evolve, optimization strategies and user-based learning approaches have been explored to enhance summary quality, diversity, and readability. These methods leverage advanced metrics and algorithms to refine both extractive and abstractive summarization models.

One notable example is the PDSum method [106], which introduces N-RL and N-BS metrics to assess lexical and semantic novelty in generated summaries. These metrics compare newly produced summaries to prior ones within the same document set, ensuring that summaries maintain diversity and avoid excessive repetition. By quantifying the distinctiveness of information, PDSum enhances the variety of content included in a summarization model, ultimately leading to a richer representation of key ideas across multiple documents.

Another approach, the EMDS framework [27], places emphasis on improving the readability of extractive summaries. This method integrates k-means clustering with the Flesch readability index to generate more comprehensible summaries. By segmenting sentences into semantically coherent clusters, EMDS ensures that selected content aligns with readability standards, surpassing conventional summarization techniques such as LSA and LDA. The integration of readability-focused evaluation metrics further enhances the accessibility and fluency of machine-generated summaries.

4. Discussion

In this section, we will discuss the answers to our research questions.

4.1. RQ1: What are the main techniques used in long document summarization, and how are they categorized?

Between 2022 and 2024, the landscape of long document summarization has experienced a notable shift toward hybrid models that integrate both extractive and abstractive techniques, frequently enhanced with deep learning and optimization strategies. These hybrid models are designed to retain the factual accuracy of extractive approaches while benefiting from the linguistic flexibility of abstractive generation.

Among the most relevant techniques are:

Hierarchical Transformers: Designed to handle long contexts through multi-level attention mechanisms. Models such as Longformer and Hierarchical BERT process documents in chunks and recombine information hierarchically.

Graph-based Models: Approaches like TOMDS and others leverage Graph Neural Networks to capture inter-sentence and discourse-level dependencies, improving coherence and contextual fidelity.

Domain-specific Pretrained Models: Variants such as BioBART, PubMedBERT, LegalBERT, and Lawformer showcase the power of transfer learning in specialized domains such as medicine and law, where maintaining domain-specific terminology and factual correctness is crucial.

Optimization-enhanced Summarizers: Models like PDSum incorporate optimization heuristics (e.g., Firefly Algorithm, Ant Colony Optimization) to guide the selection of content that maximizes coverage, diversity, and novelty (1).

Formally, many techniques define a multi-objective scoring function:

$$\text{Score} = \lambda_1 \cdot \text{Coverage}(S) + \lambda_2 \cdot \text{Diversity}(S) + \lambda_3 \cdot \text{Novelty}(S) \tag{1}$$

where S is the set of selected sentences, and the weights λ_i adjust the relative contribution of each component.

These hybrid approaches are especially successful in domains where summaries must be concise yet informative, such as biomedical literature reviews, legal case summaries, and educational material generation. Their flexibility in adapting to document structure, topic variety, and domain-specific language makes them highly promising for real-world deployment.

4.2. RQ2: What are the most commonly used evaluation metrics for assessing the quality of long document summaries in recent studies?

Evaluation metrics for long document summarization have evolved considerably, transitioning from surface-level comparisons to more semantically grounded assessments. The most prevalent metrics can be grouped as follows:

4.2.1. Lexical overlap metrics

ROUGE: Still the most cited metric for benchmarking [117], despite its known limitations in capturing paraphrasing or semantic equivalence. ROUGE metrics may fail to capture deeper semantic relationships such as paraphrasing or synonymy. As shown in Eq. (2), ROUGE-N evaluates the precision of matching n-grams, while Eq. (3) accounts for the longest common subsequence (LCS) between the generated and reference summaries.

For ROUGE-N, the formula is:

$$\text{ROUGE-N} = \frac{\sum_{n\text{-gram}} \text{match}(n\text{-gram}, R)}{\sum_{n\text{-gram}} \text{count}(n\text{-gram}, G)} \tag{2}$$

where:

- n-gram is a sequence of n words (e.g., unigrams, bigrams).
- match(n-gram, R) counts the number of matching n-grams between the generated summary and the reference.
- count(n-gram, G) counts the number of n-grams in the generated summary.

For ROUGE-L, which considers the longest common subsequence, the formula is:

$$\text{ROUGE-L} = \frac{\sum_l \text{LCS}(l, G, R)}{\text{Length of } R} \tag{3}$$

where:

- LCS(l, G, R) is the length of the longest common subsequence between the generated summary and the reference.
- Length of R is the total length of the reference summary.

BLEU: The BLEU metric often used in machine translation, evaluates the overlap of n-grams while incorporating a brevity penalty to adjust for length discrepancies between the generated and reference summaries. This metric is defined in Eq. (4) as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log P_n\right) \tag{4}$$

where:

- BP is the Brevity Penalty, which penalizes short translations.
- P_n is the precision for each n-gram (how many n-grams in the generated text match the reference).
- w_n is the weight assigned to each precision P_n.

4.2.2. Semantic similarity metrics

BERTScore: calculates semantic similarity between token embeddings using a pre-trained BERT model [118]. It reports Precision (5), Recall (6), and F1 (7) scores as follows:

$$\text{Precision} = \frac{1}{|T|} \sum_{i \in T} \max_{j \in R} \text{cosine}(e_i, e_j) \tag{5}$$

$$\text{Recall} = \frac{1}{|R|} \sum_{j \in R} \max_{i \in T} \text{cosine}(e_j, e_i) \tag{6}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Here, e_i and e_j are token embeddings from the generated summary T and the reference summary R, respectively.

MoverScore: Measures the minimal cost to transform one distribution of embeddings into another (8) using Earth Mover's Distance. It is useful for comparing summaries with significant word order variation.

$$\text{MoverScore} = \text{EMD}(T, R, \mathbf{C}) \tag{8}$$

where $T = \{t_1, t_2, \ldots, t_m\}$ and $R = \{r_1, r_2, \ldots, r_n\}$ are the tokens from the generated and reference texts, respectively, and $\mathbf{C}_{ij} = 1 - \text{cosine}(e_{t_i}, e_{r_j})$ is the cost matrix computed from cosine distances between token embeddings.

TF-IDF Cosine Similarity: Employed mainly in extractive systems to compare importance-weighted term vectors of source and summary.

4.2.3. Readability and fluency metrics

Metrics such as **Flesch-Kincaid Grade Level** (9), **Gunning Fog Index** (10), and **SMOG Index** (11) are used to assess the accessibility of generated summaries.

The *Flesch-Kincaid Grade Level* estimates the U.S. school grade level required to understand the text and is computed as:

$$\text{FKGL} = 0.39 \cdot \frac{words}{sentences} + 11.8 \cdot \frac{syllables}{words} - 15.59 \tag{9}$$

The *Gunning Fog Index* indicates the years of formal education needed to understand the text on the first reading:

$$\text{GFI} = 0.4 \cdot \left(\frac{words}{sentences} + 100 \cdot \frac{complex\ words}{words} \right) \tag{10}$$

The *SMOG Index* (Simple Measure of Gobbledygook) estimates the years of education needed based on polysyllabic word count:

$$\text{SMOG} = 1.043 \cdot \sqrt{30 \cdot \frac{polysyllables}{sentences}} + 3.1291 \tag{11}$$

These metrics are often combined in multi-metric evaluations to capture different quality aspects, including lexical matching, semantic fidelity, and readability. To better understand how these different types of metrics are adopted in practice, we analyzed their usage frequency across recent studies.

As illustrated in Figs. 5 and 6, lexical metrics such as ROUGE-1, ROUGE-2, and ROUGE-L dominate both single- and multi-document summarization settings. In contrast, semantic-based metrics like BERTScore and MoverScore appear considerably less frequently. Several factors may contribute to this disparity. First, semantic metrics often demand higher computational resources and rely on large pretrained language models, which can be prohibitive for large-scale or resource-constrained experiments. Second, the lack of standardized implementation and benchmark protocols may hinder their widespread adoption, as reproducibility becomes more challenging. The interpretability of semantic scores is often less intuitive compared to overlap-based metrics, limiting their practical appeal. These considerations partly explain the observed distribution in metric usage and underscore the need for more accessible, interpretable, and computationally efficient semantic evaluation tools.

4.3. RQ3: What advancements in deep learning have been incorporated into long document summarization?

Recent advancements in deep learning have played a pivotal role in expanding the capabilities of LDS systems. One of the key breakthroughs involves the ability to handle extended input sequences. Traditional transformer models are limited by token constraints, but newer architectures like Longformer, BigBird, and LED introduce sparse and global attention mechanisms that allow for the processing of significantly longer texts. This enhancement is crucial for preserving context and coherence in lengthy documents.

Another important development is the rise of retrieval-augmented generation (RAG) models, which integrate external information retrieval with generative modeling. By referencing relevant external documents during summarization, these models can produce outputs that are not only more informative, but also factually grounded, which is particularly valuable in knowledge-intensive tasks.

Large language models that are prompt-based or instruction-tuned, such as GPT-3.5-turbo and FLAN-T5, have also become central to modern summarization strategies. These models can perform summarization tasks directly from natural language instructions, which reduce the reliance on task-specific fine-tuning and allow for greater adaptability across domains and user needs.

In parallel, the field has seen growing interest in multimodal summarization, where textual input is enriched with non-textual data like images, tables, or videos. This is particularly relevant in areas such as education, journalism, and scientific communication, where visual context can enhance the informativeness of the summary.

Furthermore, the incorporation of human-in-the-loop strategies has introduced mechanisms for real-time feedback and correction. These approaches help mitigate issues such as hallucinations and improve the factual reliability of the output. Reinforcement learning with human feedback (RLHF) and self-supervised learning techniques are also being explored as ways to better align model behavior with human expectations, even in the absence of large labeled datasets.

4.4. RQ4: Are there any limitations identified in recent literature regarding the metrics used to evaluate summary quality?

Yes, recent literature identifies several important limitations in the metrics commonly used to evaluate summary quality. One of the most prominent issues is that traditional metrics such as ROUGE and BLEU often fail to capture semantic equivalence. They tend to penalize summaries that use synonyms or paraphrases, even when the intended meaning is preserved. This results in a semantic blind spot that limits their usefulness in evaluating more abstract or creative outputs.

These limitations are exacerbated in cross-lingual summarization, where lexical overlap is less indicative due to linguistic and cultural variation. Standard metrics fail to capture model performance on multilingual datasets like LoRaLay or CAIL-2020, especially in low-resource languages, where true improvements may go undetected without semantically-informed, language-sensitive evaluation tools.

Another significant limitation is the lack of factual validation. Current metrics are generally unable to detect hallucinated content or factual inaccuracies in summaries, an especially critical flaw in high-stakes domains such as legal or medical texts. Additionally, these metrics often show low correlation with human judgments. Studies have found that automatic scores may not align well with human evaluations of coherence, fluency, or relevance, raising concerns about their reliability.

Even newer, semantically informed metrics like BERTScore present challenges. While they offer better alignment with meaning, they rely on complex embeddings that reduce transparency and make their outputs harder to interpret. They also come with a high computational cost, which can hinder their use in large-scale or real-time applications.

In response to these limitations, recent research has begun to propose more holistic evaluation strategies. These include explainable metrics that offer interpretable scoring, tools for checking factual consistency against source documents, and models that incorporate user preferences or human feedback to better approximate human evaluation. Although these innovations represent meaningful progress, existing metrics are still insufficient for capturing the full range of qualities that define a good summary.

Thus, the literature clearly acknowledges that current evaluation methods fall short in several key areas, and there is an active movement toward the development of more comprehensive, interpretable, and context-aware evaluation frameworks.

4.5. Limitations of metrics and future directions

Despite their historical usefulness, traditional metrics such as ROUGE and BLEU exhibit significant limitations. They penalize linguistic reformulations, are insensitive to global meaning, and fail to detect factual inconsistencies. This restricts their effectiveness, particularly in domains where semantic fidelity and factual accuracy are critical, such as medicine, law, and public policy.

Meanwhile, more sophisticated metrics like BERTScore and MoverScore have advanced the evaluation of semantic similarity, but they still face key challenges. Their heavy reliance on pretrained language models introduces potential biases, and their outputs are not always intuitively interpretable. Furthermore, these metrics often require significant computational resources, making large-scale or real-time evaluations less feasible. This raises concerns about the trade-off between evaluation quality and efficiency, especially in production environments where rapid feedback is essential.

Looking ahead, we anticipate a shift towards hybrid evaluation frameworks that integrate lexical overlap, semantic understanding, and pragmatic elements (e.g., factual accuracy, coherence, and utility). These frameworks should aim for explainability, enabling developers and users to understand why a given summary is evaluated as good or poor. Additionally, there is increasing interest in self-supervised evaluation metrics that do not require reference summaries and instead learn to model human preferences or factual correctness directly.

Recent developments also aim to overcome the context length limitations of LLMs, which remain a bottleneck for long-document summarization. Approaches such as LongLLMLingua [119] leverage prompt compression to accelerate and enhance LLMs in long-context scenarios, while RAG pipelines combine external document retrieval with generative models to improve both relevance and factual grounding. These innovations reflect the ongoing evolution of summarization systems and their growing adaptability to real-world constraints.

Beyond metric development, emerging trends in summarization point toward disruptive techniques such as instruction tuning and prompting, which allow LLMs like GPT-3.5-turbo to generate summaries with minimal fine-tuning. Similarly, multimodal summarization, which combines textual input with images, video, or audio, offers rich potential for domains like education and journalism. Another promising direction is the integration of human-in-the-loop systems, where users can guide or correct summaries during generation, enhancing both quality and trust.

While this review primarily focuses on textual long-document summarization, we acknowledge the growing interest in multimodal LDS, especially in handling complex structured documents such as image-text layouts, web pages, and richly formatted PDFs. Although our systematic review identified limited coverage of this direction, most notably, the LoRaLay dataset, which introduces a multilingual and multimodal benchmark for layout-aware summarization, this remains an underexplored area in current literature. Summarization of structured content like tables and charts poses unique challenges related to spatial reasoning, information hierarchy, and modality fusion. We consider this an important avenue for future work and encourage more comprehensive reviews once the field matures further.

Future research must address the ethical and social dimensions of summarization. This includes developing metrics and techniques that are robust to bias, ensure factual reliability, and offer transparency in high-stakes contexts. Involving end-users in the evaluation process especially those impacted by automated decisions, can help center real-world utility, accessibility, and fairness in summarization research.

4.6. Integrated outlook and final remarks

The period between 2022 and 2024 has marked a turning point in the evolution of long-document summarization, with significant advancements in both modeling techniques and evaluation strategies. The emergence of hybrid summarization models, combining extractive and abstractive approaches with optimization algorithms and deep learning architectures, has enabled the generation of more coherent, contextually aware, and semantically rich summaries. At the same time, the rise of semantic evaluation metrics, such as BERTScore and MoverScore, has provided a more nuanced assessment of summary quality, addressing some of the limitations of traditional n-gram overlap metrics.

However, these developments remain constrained by persistent challenges: the inability to process very long contexts efficiently, the prevalence of hallucinations, the lack of interpretability, and the misalignment between automatic metrics and human judgment. Addressing these issues requires not only technical innovation, but also a shift in evaluation philosophy, one that prioritizes semantic fidelity, human-centered assessments, and domain adaptability.

Moving forward, we foresee a co-evolution between summarization models and evaluation metrics, where improvements in one domain necessitate advancements in the other. This includes the development of efficient and transparent models capable of handling complex documents, as well as metrics that better capture factual accuracy, coherence, and practical utility. Moreover, new paradigms, such as instruction-tuned LLMs, multimodal summarization, and human-in-the-loop frameworks, are expected to reshape the landscape, bringing summarization systems closer to real-world applicability.

Ultimately, the future of summarization lies in bridging the gap between algorithmic sophistication and human needs. Emphasizing ethical design, computational efficiency, and contextual relevance will be key to building robust summarization systems that are not only technically sound, but also socially responsible and truly useful in diverse domains.

5. Conclusion

The study highlights that hybrid models, which blend extractive and abstractive techniques, have emerged as the dominant approach in long document summarization. These models often leverage optimization algorithms such as Firefly and Ant Colony Optimization alongside advanced deep learning architectures like BERT, Longformer, and Graph Neural Networks to enhance the informativeness, coherence, and diversity of summaries. In domain-specific contexts like law and medicine, specialized models such as LegalBERT and BioBERT underscore the value of tailored solutions.

In terms of evaluation, there has been a noticeable shift from traditional surface-level n-gram metrics like ROUGE and BLEU toward more semantically aware measures, including BERTScore and MoverScore. While these newer metrics offer improved meaning alignment, they also introduce challenges related to interpretability, bias, and computational demands. Complementary tools such as Flesch-Kincaid scores and human evaluations continue to be essential for assessing readability and factual accuracy.

Recent advancements in deep learning have further shaped the field, with innovations such as models capable of handling longer contexts, instruction-tuned large language models like FLAN-T5 and GPT-3.5, and RAG frameworks. Reinforcement learning and human-in-the-loop strategies have also played a critical role in boosting the reliability and factual grounding of generated summaries.

Nonetheless, persistent challenges remain, particularly in the evaluation process. Traditional metrics often overlook issues like hallucinations or semantic drift, while more advanced approaches bring their own trade-offs in transparency and efficiency. As a result, there is growing interest in multi-dimensional evaluation frameworks that balance semantic similarity, readability, factual consistency, and user-centered perspectives.

While the field has made significant progress, this review is constrained by its focus on English-language and indexed sources. Future work should prioritize the development of interpretable, efficient evaluation techniques, especially for specialized domains. Equally important is the integration of ethical safeguards, mechanisms for bias detection,

and avenues for user participation to ensure that LDS systems are not only effective, but also fair and socially responsible.

CRediT authorship contribution statement

Bady Gana: Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Héctor Allende-Cid:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis, Conceptualization. **Stefan Rüping:** Writing – review & editing, Supervision, Resources, Project administration, Investigation. **Marcelo Becerra-Rozas:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization. **Juan Zamora:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Bady Gana is supported by National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO NACIONAL/2024–21240115. Bady Gana is supported by Beca INF-PUCV. Marcelo Becerra-Rozas is supported by DI Iniciación PUCV 2025/PUCV/039.725/2025. Juan Zamora is supported by Grant ANID/FONDECYT/INICIACION/11200826.

Data availability

Data will be made available upon request.

References

[1] H. Shakil, A. Farooq, J. Kalita, Abstractive text summarization: state of the art, challenges, and improvements, Neurocomputing 603 (2024) 128255, https://doi.org/10.1016/j.neucom.2024.128255.
[2] H.Y. Koh, J. Ju, M. Liu, S. Pan, An empirical survey on long document summarization: datasets, models, and metrics, ACM Comput. Surv. 55 (8) (2023) 1–35, https://doi.org/10.1145/3545176.
[3] M.H.H. Wahab, N.H. Ali, N.A.W. Abdul Hamid, S.K. Subramaniam, R. Latip, M. Othman, A review on optimization-based automatic text summarization approach, IEEE Access 12 (2024) 4892–4909, https://doi.org/10.1109/ACCESS.2023.3348075.
[4] A. Elsaid, A. Mohammed, L.F. Ibrahim, M.M. Sakre, A comprehensive review of arabic text summarization, IEEE Access 10 (2022) 38012–38030, https://doi.org/10.1109/ACCESS.2022.3163292.
[5] A. Chaves, C. Kesiku, B. Garcia-Zapirain, Automatic text summarization of biomedical text data: a systematic review, Information 13 (8) (2022) 393, https://doi.org/10.3390/info13080393.
[6] N. Giarelis, C. Mastrokostas, N. Karacapilidis, Abstractive vs. extractive summarization: an experimental review, Appl. Sci. 13 (13) (2023) 7620, https://doi.org/10.3390/app13137620.
[7] A. Mishra, M. Naruka, S. Tiwari, Extraction techniques and evaluation measures for extractive text summarisation, 2023, https://doi.org/10.1007/978-3-031-13577-4_17, pp. 279–290.
[8] E. Alanzi, S. Alballaa, Query-focused multi-document summarization survey, Int. J. Adv. Comput. Sci. Appl. 14 (6) (2023) https://doi.org/10.14569/IJACSA.2023.0140688.
[9] C. Meaney, T.A. Stukel, P.C. Austin, R. Moineddin, M. Greiver, M. Escobar, Quality indices for topic model selection and evaluation: a literature review and case study, BMC Med. Inform. Decis. Mak. 23 (1) (2023) 132, https://doi.org/10.1186/s12911-023-02216-1.
[10] M. Barbella, M. Risi, G. Tortora, A. Auriemma Citarella, Different metrics results in text summarization approaches:, in: Proceedings of the 11th International Conference on Data Science, Technology and Applications, SCITEPRESS - Science and Technology Publications, Lisbon, Portugal, 2022, pp. 31–39, https://doi.org/10.5220/0011144000003269.
[11] A. Auriemma Citarella, M. Barbella, M.G. Ciobanu, F. De Marco, L. Di Biasi, G. Tortora, Assessing the effectiveness of ROUGE as unbiased metric in extractive vs. abstractive summarization techniques, J. Comput. Sci. 87 (2025) 102571, https://doi.org/10.1016/j.jocs.2025.102571, https://www.sciencedirect.com/science/article/pii/S1877750325000481.
[12] M. Afsharizadeh, H. Ebrhimpour-Komeleh, A. Bagheri, G. Chrupała, A survey on multi-document summarization and domain-oriented approaches, J. Inf. Syst. Telecommun. 10 (37) (2022) 68–78, https://doi.org/10.52547/jist.16245.10.37.68.
[13] D. Aumiller, J. Fan, M. Gertz, On THE state of German (abstractive) text summarization, Gesellschaft für Informatik e.V., ISBN: 9783885797258, 2023, https://doi.org/10.18420/BTW2023-10.
[14] K. Hewapathirana, N. De Silva, C.D. Athuraliya, Multi-document summarization: a comparative evaluation, in: 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), IEEE, Peradeniya, Sri Lanka, 2023, pp. 19–24, https://doi.org/10.1109/ICIIS58898.2023.10253581.
[15] E. Davoodijam, M. Alambardar Meybodi, Evaluation metrics on text summarization: comprehensive survey, Knowl. Inf. Syst. 66 (2024) 7717–7738, https://doi.org/10.1007/s10115-024-02217-0.
[16] M.J. Page, D. Moher, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, J.E. McKenzie, Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, BMJ 372 (2021) https://doi.org/10.1136/bmj.n160.
[17] B. Mutlu, E. Sezer, Enhanced sentence representation for extractive text summarization: investigating the syntactic and semantic features and their contribution to sentence scoring, Expert Syst. Appl. 227 (2023) 120302, https://doi.org/10.1016/j.eswa.2023.120302.
[18] B. Ma, Mining both commonality and specificity from multiple documents for multi-document summarization, IEEE Access (2024) 54371–54381, https://doi.org/10.1109/ACCESS.2024.3388493.
[19] P. Mahalakshmi, D. Fatima, Summarization of text and image captioning in information retrieval using deep learning techniques, IEEE Access 10 (2022) 18289–18297, https://doi.org/10.1109/ACCESS.2022.3150414.
[20] P. Patel, P. Solanki, An impact of Firefly multi-objective optimization algorithm in the process of text summarization for generation good summaries, J. Integr. Sci. Technol. 12 (May 2024) https://doi.org/10.62110/sciencein.jist.2024.v12.834.
[21] S. Mulla, N. Shaikh, Effective Elytron Vespid-b rank BILSTM classifier for multi-document summarization, Multimed. Tools Appl. 83 (2023) 1–22, https://doi.org/10.1007/s11042-023-17544-7.
[22] H. Zhang, S. Cho, K. Song, X. Wang, H. Wang, Z. Jiawei, D. Yu, Unsupervised multi-document summarization with holistic inference, (2023) 123–133, https://doi.org/10.18653/v1/2023.findings-ijcnlp.11.
[23] S. Lamsiyah, A. Mahdaouy, S. Ouatik El Alaoui, B. Espinasse, Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, BM25 model, and maximal marginal relevance criterion, J. Ambient Intell. Humaniz. Comput. 14 (Mar 2021) https://doi.org/10.1007/s12652-021-03165-1.
[24] M. Murali Krishna, J. Vankara, S. Nandini, G. Karetla, K. Naidu, Multi-objective ant colony optimization (MOACO) approach for multi-document text summarization, Eng. Proc. (2024) 218, https://doi.org/10.3390/engproc2023059218.
[25] J. Tamilselvan, A. Senthilrajan, A novel approach for multi-document summarization to produce readable text using k-means and Flesch score, Indian J. Comput. Sci. Eng. 13 (2022) 238–247, https://doi.org/10.21817/indjcse/2022/v13i1/221301161.
[26] X. Sun, H. Zhuge, Generating survey draft based on closeness of position distributions of key words, Expert Syst. Appl. 237 (2023) 121422, https://doi.org/10.1016/j.eswa.2023.121422.
[27] M. Turan, Selection informative units for extractive summarization, WSEAS Trans. Syst. 22 (2023) 287–294, https://doi.org/10.37394/23202.2023.22.31.
[28] R. Rani, D.K. Lobiyal, Document vector embedding based extractive text summarization system for Hindi and English text, Appl. Intell. 52 (8) (2022) 9353–9372, https://doi.org/10.1007/s10489-021-02871-9.
[29] R. Bandaru, Y. Radhika, Extractive multi-document text summarization leveraging hybrid semantic similarity measures, Int. J. Adv. Comput. Sci. Appl. 13 (Jan 2022) https://doi.org/10.14569/IJACSA.2022.0130998.
[30] J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, C. Pérez, A multi-objective memetic algorithm for query-oriented text summarization: medicine texts as a case study, Expert Syst. Appl. 198 (2022) 116769, https://doi.org/10.1016/j.eswa.2022.116769.
[31] J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, C. Pérez, Automatic update summarization by a multiobjective number-one-selection genetic approach, IEEE Trans. Cybern. (Dec 2022) https://doi.org/10.1109/TCYB.2022.3223163.
[32] J.M. Sanchez-Gomez, M.A. Vega-Rodríguez, C. Pérez, An indicator-based multi-objective variable neighborhood search approach for query-focused summarization, Swarm Evol. Comput. 91 (2024) 101721, https://doi.org/10.1016/j.swevo.2024.101721.
[33] A. Ghadimi, H. Beigy, SGCSUMM: an extractive multi-document summarization method based on pre-trained language model, submodularity, and graph convolutional neural networks, Expert Syst. Appl. 215 (2023) 119308, https://doi.org/10.1016/j.eswa.2022.119308.
[34] P. Pitchandi, Document clustering using graph based fuzzy association rule generation, Comput. Syst. Sci. Eng. 43 (2022) 203–218, https://doi.org/10.32604/csse.2022.020459.
[35] T. Vo, A novel semantic-enhanced generative adversarial network for abstractive text summarization, Soft Comput. 27 (2023) 1–14, https://doi.org/10.1007/s00500-023-07890-x.
[36] M. Han, Z. Wang, H. Wang, X. Bao, K. Niu, Abstractive multi-document summarization with cross-documents discourse relations, 2023, https://doi.org/10.1007/978-981-99-8145-8_36, pp. 470–481.

[37] S. Mishra, N. Saini, S. Saha, P. Bhattacharyya, Scientific document summarization in multi-objective clustering framework, Appl. Intell. 52 (2022) 1–24, https://doi.org/10.1007/s10489-021-02376-5.
[38] A. Alambo, T. Banerjee, K. Thirunarayan, M. Raymer, Entity-driven fact-aware abstractive summarization of biomedical literature, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, Montreal, QC, Canada, 2022, pp. 613–620, https://doi.org/10.1109/ICPR56361.2022.9956656.
[39] S. Bano, S. Khalid, N.M. Tairan, H. Shah, H.A. Khattak, Summarization of scholarly articles using BERT and BIGRU: deep learning-based extractive approach, J. King Saud Univ. Comput. Inf. Sci. 35 (9) (2023) 101739. https://doi.org/10.1016/j.jksuci.2023.101739.
[40] A. Khaliq, A. Khan, S. Afsar Awan, S. Jan, M. Umair, M.F. Zuhairi, Integrating topic-aware heterogeneous graph neural network with transformer model for medical scientific document abstractive summarization, IEEE Access 12 (2024) 113855–113866, https://doi.org/10.1109/ACCESS.2024.3443730.
[41] D. Jain, M.D. Borah, A. Biswas, Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization, Knowl.-Based Syst. 264 (2023) 110336, https://doi.org/10.1016/j.knosys.2023.110336.
[42] D. Jain, M.D. Borah, A. Biswas, Summarization of lengthy legal documents via abstractive dataset building: an extract-then-assign approach, Expert Syst. Appl. 237 (2024) 121571, https://doi.org/10.1016/j.eswa.2023.121571.
[43] D. Jain, M.D. Borah, A. Biswas, A sentence is known by the company it keeps: improving legal document summarization using deep clustering, Artif. Intell. Law 32 (1) (2024) 165–200. https://doi.org/10.1007/s10506-023-09345-y.
[44] H. Mahmoud, A. Hafez, A novel optimized language-independent text summarization technique, Comput. Mater. Contin. 73 (2022) 5121–5136, https://doi.org/10.32604/cmc.2022.031485.
[45] Q. Wang, R. Wang, K. Zhao, R. Amor, B. Liu, X. Zheng, Z. Zhang, Z. Huang, Towards legal judgment summarization: a structure-enhanced approach, 2023, https://doi.org/10.3233/FAIA230553.
[46] S. Yu, W. Gao, Y. Qin, C. Yang, R. Huang, Y. Chen, C. Lin, ITERSUM: iterative summarization based on document topological structure, Inf. Process. Manag. 62 (2025) 103918, https://doi.org/10.1016/j.ipm.2024.103918.
[47] J. Ouyang, W. Huang, T. Liu, A hybrid automatic text summarization model for judgment documents, (2022) 374–386, https://doi.org/10.1007/978-3-031-06767-9_31.
[48] X. Shen, W. Lam, S. Ma, H. Wang, Joint learning of text alignment and abstractive summarization for long documents via unbalanced optimal transport, Nat. Lang. Eng. 30 (3) (2024) 525–553, https://doi.org/10.1017/S1351324923000177.
[49] C. Han, J. Feng, H. Qi, Topic model for long document extractive summarization with sentence-level features and dynamic memory unit, Expert Syst. Appl. 238 (2023) 121873, https://doi.org/10.1016/j.eswa.2023.121873.
[50] A. Ghadimi, H. Beigy, Hybrid multi-document summarization using pre-trained language models, Expert Syst. Appl. 192 (2021) 116292, https://doi.org/10.1016/j.eswa.2021.116292.
[51] W. Su, J. Jiang, K. Huang, Multi-granularity adaptive extractive document summarization with heterogeneous graph neural networks, PeerJ Comput. Sci. 9 (2023) e1737, https://doi.org/10.7717/peerj-cs.1737.
[52] H. Bakr, S. Mohamed, Automatic multi-documents text summarization by a large-scale sparse multi-objective optimization algorithm, Complex Intell. Syst. 9 (Sep 2023) https://doi.org/10.1007/s40747-023-00967-y.
[53] M. Tomer, M. Kumar, Multi-document extractive text summarization based on Firefly algorithm, J. King Saud Univ. Comput. Inf. Sci. 34 (Apr 2021) https://doi.org/10.1016/j.jksuci.2021.04.004.
[54] A. Arora, Ensemble of support vector machine and ontological structures to generate abstractive text summarization, Int. J. Inf. Retr. Res. 12 (2022) 1–24, https://doi.org/10.4018/IJIRR.300294.
[55] M. Jain, R. Jindal, A. Jain, An experimental study of game theory with various word embeddings for automatic extractive text summarization, Multimed. Tools Appl. (2024) 1–23, https://doi.org/10.1007/s11042-024-19828-y.
[56] B. Mohamed, S. Aitouche, H. Moumen, H. Houassi, H. Rahab, B. Abdelaali, Bbmamds: binary biology migration algorithm for multi-document text summarization, Rev. Intell. Artif. 37 (2023) 1147–1158, https://doi.org/10.18280/ria.370506.
[57] S. Patil, R. Rautray, SMATS: single and multi automatic text summarization, Karbala Int. J. Mod. Sci. 9 (Jan 2023) https://doi.org/10.33640/2405-609X.3281.
[58] H. Aliakbarpour, M.T. Manzuri, A.M. Rahmani, Automatic text summarization using deep reinforced model coupling contextualized word representation and attention mechanism, Multimed. Tools Appl. 83 (1) (2024) 733–762. https://doi.org/10.1007/s11042-023-15589-2.
[59] Z. Jalil, M. Nasir, M. Alazab, J. Nasir, T. Amjad, A. Alqammaz, Grapharizer: a graph-based technique for extractive multi-document summarization, Electronics 12 (8) (2023) 1895. https://doi.org/10.3390/electronics12081895.
[60] R. Malik, K. Khan, W. Nawaz, Maximal Gspan: multi-document summarization through frequent subgraph mining, 2023, https://doi.org/10.1109/IMCOM56909.2023.10035618, pp. 1–7.
[61] A. Karotia, S. Susan, Pre-training meets clustering: a hybrid extractive multi-document summarization model, 2023, https://doi.org/10.1007/978-3-031-27409-1_48, pp. 532–542.
[62] M.T.R. Laskar, E. Hoque, J.X. Huang, Domain adaptation with pre-trained transformers for query-focused abstractive text summarization, Comput. Linguist. 48 (2) (2022) 279–320. https://doi.org/10.1162/coli_a_00434.
[63] S. Lamsiyah, A.E. Mahdaouy, C. Schommer, Can anaphora resolution improve extractive query-focused multi-document summarization?, IEEE Access 11 (2023) 99961–99976, https://doi.org/10.1109/ACCESS.2023.3314524.
[64] M. Mohamed, M. Oussalah, V. Chang, Sdbqfsum: query-focused summarization framework based on diversity and text semantic analysis, Expert Syst. 41 (Sep 2023) https://doi.org/10.1111/exsy.13462.
[65] S. Gong, Z. Zhu, J. Qi, W. Wu, C. Tong, Sebursum: a novel set-based summary ranking strategy for summary-level extractive summarization, J. Supercomput. 79 (12) (2023) 12949–12977, https://doi.org/10.1007/s11227-023-05165-8.
[66] H. Zhao, W. Zhang, M. Huang, S. Feng, Y. Wu, A multi-granularity heterogeneous graph for extractive text summarization, Electronics 12 (2023) 2184, https://doi.org/10.3390/electronics12102184.
[67] T. Wang, C. Yang, M. Zou, J. Liang, D. Xiang, W. Yang, H. Wang, J. Li, A study of extractive summarization of long documents incorporating local topic and hierarchical information, Sci. Rep. 14 (May 2024) https://doi.org/10.1038/s41598-024-60779-z.
[68] S. Yang, S. Zhang, M. Fang, F. Yang, S. Liu, A hierarchical representation model based on longformer and transformer for extractive summarization, Electronics 11 (2022) 1706, https://doi.org/10.3390/electronics11111706.
[69] G. Swetha, S. Kumar, A hierarchical framework based on transformer technology to achieve factual consistent and non-redundant abstractive text summarization, Multimed. Tools Appl. 83 (Oct 2023) https://doi.org/10.1007/s11042-023-17426-y.
[70] P. Kherwa, J. Arora, T. Sharma, G. Deepali, S. Juneja, G. Muhammad, A. Nauman P.hD, Contextual embedded text summarizer system: a hybrid approach, Expert Syst. 42 (Oct 2024) https://doi.org/10.1111/exsy.13733.
[71] D. Qiu, B. Yang, Text summarization based on multi-head self-attention mechanism and pointer network, Complex Intell. Syst. 8 (Sep 2021) https://doi.org/10.1007/s40747-021-00527-2.
[72] C.D. Chelliah, S. Parthasarathy, Ext-icas: a novel self-normalized extractive intra cosine attention similarity summarization, Comput. Syst. Sci. Eng. 45 (2023) 377–393, https://doi.org/10.32604/csse.2023.027481.
[73] A. Dalal, S. Ranjan, Y. Bopaiah, D. Chembachere, N. Steiger, C. Burns, V. Daswani, Text summarization for pharmaceutical sciences using hierarchical clustering with a weighted evaluation methodology, Sci. Rep. 14 (Aug 2024) https://doi.org/10.1038/s41598-024-70618-w.
[74] J. Bian, X. Huang, H. Zhou, T. Huang, S. Zhu, GOSUM: extractive summarization of long documents by reinforcement learning and graph-organized discourse state, Knowl. Inf. Syst. 66 (2024) 7557–7580, https://doi.org/10.1007/s10115-024-02195-3.
[75] D. Onah, E. Pang, M. El-Haj, A data-driven latent semantic analysis for automatic text summarization using lda topic modelling, 2022, https://doi.org/10.1109/BigData55660.2022.10020259.
[76] M. Ulker, A. Ozer, Abstractive summarization model for summarizing scientific article, IEEE Access (2024) 91252–91262, https://doi.org/10.1109/ACCESS.2024.3420163.
[77] D. Fitrianah, R. Jauhari, Extractive text summarization for scientific journal articles using long short-term memory and gated recurrent units, Bull. Electr. Eng. Inform. 11 (2022) 150–157, https://doi.org/10.11591/eei.v11i1.3278.
[78] X. Zhang, Q. Wei, Q. Song, P. Zhang, TOMDS (topic-oriented multi-document summarization): enabling personalized customization of multi-document summaries, Appl. Sci. 14 (2024) 1880, https://doi.org/10.3390/app14051880.
[79] J.P. Verma, S. Bhargav, M. Bhavsar, P. Bhattacharya, A. Bostani, S. Chowdhury, J. Webber, A. Mehbodniya, Graph-based extractive text summarization sentence scoring scheme for big data applications, Information 14 (9) (2023) 472, https://doi.org/10.3390/info14090472.
[80] P. Wang, S. Li, J. Tang, T. Wang, What can rhetoric bring us? Incorporating rhetorical structure into neural related work generation, Expert Syst. Appl. 251 (2024) 123781, https://doi.org/10.1016/j.eswa.2024.123781.
[81] P. Wang, S. Li, S. Liu, J. Tang, T. Wang, Plan and generate: explicit and implicit variational augmentation for multi-document summarization of scientific articles, Inf. Process. Manag. 60 (4) (2023) 103409, https://doi.org/10.1016/j.ipm.2023.103409.
[82] P.P.S. Bedi, M. Bala, K. Sharma, Extractive summarization using concept-space and keyword phrase, Expert Syst. 39 (10) (2022) e13110, https://doi.org/10.1111/exsy.13110.
[83] T.S. Barros, C.E.S. Pires, D.C. Nascimento, Leveraging BERT for extractive text summarization on federal police documents, Knowl. Inf. Syst. 65 (11) (2023) 4873–4903, https://doi.org/10.1007/s10115-023-01912-8.
[84] S. Liu, J. Cao, Z. Deng, W. Zhao, R. Yang, Z. Wen, P. Yu, Neural abstractive summarization for long text and multiple tables, IEEE Trans. Knowl. Data Eng. (2023) 1–14, https://doi.org/10.1109/TKDE.2023.3324012.
[85] C. Arya, M. Diwakar, P. Singh, V. Singh, S. Kadry, J. Kim, Multi-document news web page summarization using content extraction and lexical chain based key phrase extraction, Mathematics 11 (2023) 1762, https://doi.org/10.3390/math11081762.
[86] A. Khan, F. Al-Obeidat, A. Khalid, A. Amin, F. Moreira, Sentence embedding approach using LSTM auto-encoder for discussion threads summarization, Comput. Sci. Inf. Syst. 20 (2023) 1367–1387, https://doi.org/10.2298/CSIS221210055K.
[87] M. Pan, T. Li, Y. Liu, Q. Pei, E. Huang, J. Huang, A semantically enhanced text retrieval framework with abstractive summarization, Comput. Intell. 40 (Sep 2023) https://doi.org/10.1111/coin.12603.
[88] J. Dan, W. Hu, Y. Wang, Enhancing legal judgment summarization with integrated semantic and structural information, Artif. Intell. Law (2023) 1–22, https://doi.org/10.1007/s10506-023-09381-8.
[89] K. Ando, M. Komachi, T. Okumura, H. Horiguchi, Y. Matsumoto, Is in-hospital meta-information useful for abstractive discharge summary generation? (2022) 143–148, https://doi.org/10.1109/TAAI57707.2022.00034.

[90] J. Woodring, K. Perez, A. Ali-Gombe, Enhancing privacy policy comprehension through privacify: a user-centric approach using advanced language models, Comput. Secur. 145 (2024) 103997, https://doi.org/10.1016/j.cose.2024.103997.
[91] A. Deroy, K. Ghosh, S. Ghosh, Ensemble methods for improving extractive summarization of legal case judgements, Artif. Intell. Law 32 (2023) 1–59, https://doi.org/10.1007/s10506-023-09349-8.
[92] E. Monir, A. Salah, Aratsum: arabic Twitter trend summarization using topic analysis and extractive algorithms, Int. J. Comput. Intell. Syst. 17 (Sep 2024) https://doi.org/10.1007/s44196-024-00546-0.
[93] S. Khatuya, K. Sinha, N. Ganguly, S. Ghosh, P. Goyal, Instruction-guided bullet point summarization of long financial earnings call transcripts (2024) 2477–2481, https://doi.org/10.1145/3626772.3657948.
[94] S. Alrumiah, A. Al-Shargabi, Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement, Comput. Mater. Continua 70 (2021) 6205–6221, https://doi.org/10.32604/cmc.2022.021780.
[95] V. Vaissnave, P. Deepalakshmi, Modeling of automated glowworm swarm optimization based deep learning model for legal text summarization, Multimed. Tools Appl. 82 (Nov 2022) https://doi.org/10.1007/s11042-022-14171-6.
[96] C. Chootong, T. Shih, Tech-talk-sum: fine-tuning extractive summarization and enhancing BERT text contextualization for technological talk videos, Multimed. Tools Appl. 81 (Sep 2022) https://doi.org/10.1007/s11042-022-12812-4.
[97] V. Priya, V. Praveena, L.R. Sujithra, A parallel optimization and transfer learning approach for summarization in electrical power systems, Automatika 64 (2023) 1225–1233, https://doi.org/10.1080/00051144.2023.2254975.
[98] D. Feijo, V. Moreira, Improving abstractive summarization of legal rulings through textual entailment, Artif. Intell. Law 31 (Nov 2021) https://doi.org/10.1007/s10506-021-09305-4.
[99] B.M. Gurusamy, P.K. Rangarajan, S. Parathasarathy, S. Aravind, K. Hanish, G. Pavithria, Text summarization for big data analytics: a comprehensive review of GPT 2 and BERT approaches, (2023) 247–264, https://doi.org/10.1007/978-3-031-33808-3_14.
[100] W. Cai, Z. Hu, Y. Luo, D. Liang, Y. Feng, J. Chen, Multi-layer contextual passage term embedding for ad-hoc retrieval, Information 13 (2022) 221, https://doi.org/10.3390/info13050221.
[101] C. Meyer, D. Adkins, K. Pal, R. Galici, A. Garcia-Agundez, C. Eickhoff, Neural text generation in regulatory medical writing, Front. Pharmacol. 14 (2023) 1086913, https://doi.org/10.3389/fphar.2023.1086913.
[102] Y. Shi, P. Ren, J. Wang, B. Han, T. Valizadehaslani, F. Agbavor, Y. Zhang, M. Hu, L. Zhao, H. Liang, Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting, J. Biomed. Inform. 148 (2023) 104533, https://doi.org/10.1016/j.jbi.2023.104533.
[103] R. Albeer, H. Alshahad, H.J. Aleqabie, N. Al-Shakarchy, Automatic summarization of Youtube video transcription text using term frequency-inverse document frequency, Indones. J. Electr. Eng. Comput. Sci. 26 (2022) 1512, https://doi.org/10.11591/ijeecs.v26.i3.pp1512-1519.
[104] I. Al-Hussaini, D. An, A. Lee, S. Bi, C. Mitchell, Ccs explorer: relevance prediction, extractive summarization, and named entity recognition from clinical cohort studies, (2022) 5173–5181, https://doi.org/10.1109/BigData55660.2022.10020807.
[105] I. Tampe Palma, M. Mendoza, E. Milios, Neural abstractive unsupervised summarization of online news discussions (2021) 822–841, https://doi.org/10.1007/978-3-030-82196-8_60.
[106] S. Yoon, H. Chan, J. Han, Pdsum: prototype-driven continuous summarization of evolving multi-document sets stream (2023) 1650–1661, https://doi.org/10.1145/3543507.3583371.
[107] F. Bayatmakou, A. Mohebi, A. Ahmadi, An interactive query-based approach for summarizing scientific documents, Inf. Discov. Deliv. ahead-of-print (Jun 2021) https://doi.org/10.1108/IDD-10-2020-0124.
[108] G. Sharma, D. Sharma, Improving extractive text summarization performance using enhanced feature based RBM method, Rev. Intell. Artif. 36 (2022) 777–784, https://doi.org/10.18280/ria.360516.
[109] L. Nguyen, T. Scialom, B. Piwowarski, J. Staiano, Loralay: a multilingual and multimodal dataset for long range and layout-aware summarization, 2023, pp. 636–651, https://doi.org/10.18653/v1/2023.eacl-main.46.
[110] I. Al-Hussaini, A. Wu, C. Mitchell, Pathology dynamics at Biolaysumm: the trade-off between readability, relevance, and factuality in lay summarization, (2023) 592–601, https://doi.org/10.18653/v1/2023.bionlp-1.63.
[111] I. Sonata, Y. Heryadi, Embracing text summarization IN education using transformer model, ICIC Express Lett. 18 (2024) 193–200, https://doi.org/10.24507/icicel.18.02.193.
[112] S. Liu, J. Cao, Y. Li, R. Yang, Z. Wen, Low-resource court judgment summarization for common law systems, Inf. Process. Manag. 61 (2024) 103796, https://doi.org/10.1016/j.ipm.2024.103796.
[113] G. Wicaksono, M. Hakim, N. Hayatin, N. Hidayah, T. Sari, Text summarization on verdicts of industrial relations disputes using the cross-latent semantic analysis and long short-term memory, J.O.I.V.: Int. J. Inf. Vis. 7 (2023) 847–853, https://doi.org/10.30630/joiv.7.3.2052.
[114] A. Leiva-Araos, B. Gana, H. Allende-Cid, J. García, M.J. Saikia, Large scale summarization using ensemble prompts and in context learning approaches, Sci. Rep. 15 (1) (2025) 10259, https://doi.org/10.1038/s41598-025-94551-8.
[115] B. Gana, A. Leiva-Araos, H. Allende-Cid, J. García, Leveraging LLMS for efficient topic reviews, Appl. Sci. 14 (17) (2024) https://doi.org/10.3390/app14177675, https://www.mdpi.com/2076-3417/14/17/7675.
[116] V.C. Hartman, S.S. Bapat, M.G. Weiner, B.B. Navi, E.T. Sholle, T.R. Campion, A method to automate the discharge summary hospital course for neurology patients, J. Am. Med. Inform. Assoc. 30 (12) (2023) 1995–2003, https://doi.org/10.1093/jamia/ocad177.
[117] C.-Y. Lin, Rouge: a package for automatic evaluation of summaries, (2004), p. 10.
[118] T. Zhang, V. Kishore, F. Wu, K. Weinberger, Y. Artzi, Bertscore: evaluating text generation with BERT (Apr 2019) https://doi.org/10.48550/arXiv.1904.09675.
[119] H. Jiang, Q. Wu, X. Luo, D. Li, C.-Y. Lin, Y. Yang, L. Qiu, Longllmlingua: accelerating and enhancing llms in long context scenarios via prompt compression, arXiv preprint arXiv:2310.06839 (2024) https://doi.org/10.48550/arXiv.2310.06839.

Author biography

Bady Gana received his M.Sc. degree in Computer Engineering from the Pontificia Universidad Católica de Valparaíso (PUCV), Chile, in 2023. He is currently affiliated with PUCV, where his research focuses on natural language processing, machine learning, large-scale automatic summarization, and knowledge management.

Héctor Allende-Cid received his Ph.D. in Computer Engineering from Universidad Técnica Federico Santa María, Chile, in 2015. He also holds an M.Sc. in Computer Engineering. He is currently professor at the Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Chile. His research focuses on machine learning, natural language processing, and semantic technologies, with applications in clinical text mining, software analytics, and cybersecurity.

Stefan Rüping currently leads the "Knowledge Discovery" department at Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS). He holds a Ph.D. in Computer Science from the University of Dortmund (2006). Dr. Rüping has extensive experience in research, consulting, and project management bridging academia and industry. His research focuses on machine learning, artificial intelligence, knowledge discovery, and data mining, with a strong emphasis on applications in biomedical informatics, personalized medicine, and healthcare. Dr. Rüping has authored numerous influential publications in international conferences and specialized journals.

Marcelo Becerra-Rozas received his B.Sc. degree in Computer Science Engineering in 2021, his M.Sc. degree in Computer Science Engineering in 2022, and his Ph.D. in Computer Science Engineering in 2024, all from the Pontificia Universidad Católica de Valparaíso (PUCV), Chile. In 2025, he was appointed as a lecturer at the Escuela de Ingeniería Informática at PUCV. His research interests include artificial intelligence, constraint programming, metaheuristics, and combinatorial optimization.

Juan Zamora received his Ph.D., M.Sc., and professional degree in Computer Engineering from Universidad Técnica Federico Santa María, Chile. He is currently professor at the Instituto de Estadística, Pontificia Universidad Católica de Valparaíso. His research interests lie at the intersection of text mining, high-dimensional data clustering, and unsupervised learning on distributed document collections. His recent work explores large-scale knowledge extraction from unstructured data, document representation learning, and statistical modeling of textual complexity in educational and clinical contexts.