IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 September 2023, accepted 27 September 2023, date of publication 5 October 2023,
date of current version 11 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322188

SURVEY

Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey

BILAL KHAN[1], ZOHAIB ALI SHAH[1], MUHAMMAD USMAN[2], (Senior Member, IEEE), INAYAT KHAN[2], AND BADAM NIAZI[3]

[1]Department of Computer Software Engineering, University of Engineering and Technology Mardan, Mardan 23200, Pakistan
[2]Department of Computer Science, University of Engineering and Technology Mardan, Mardan 23200, Pakistan
[3]Faculty of Computer Science, Nangarhar University, Jalalabad 2601, Afghanistan

Corresponding author: Badam Niazi (badam@nu.edu.af)

ABSTRACT The discipline of Automatic Text Summarization (ATS), which is expanding quickly, intends to automatically create summaries of enormous amounts of text so that readers can save time and effort. ATS is a rapidly growing field that aims to save readers time and effort by automatically generating summaries of large volumes of text. In recent years, significant advancements have been witnessed in this area, accompanied by challenges that have spurred extensive research. The proliferation of textual data has sparked substantial interest in ATS, which is thoroughly examined in this survey study. Researchers have been refining ATS techniques since the 1950s, primarily categorized as extractive, abstractive, or hybrid approaches. In the extractive approach, key sentences are extracted from the source document(s) and combined to form the summary, while the abstractive approach employs an intermediary representation of the input document(s) to generate a summary that may differ from the original text. Hybrid approaches combine elements of both extractive and abstractive methods. Despite various recommended methodologies, the generated summaries still exhibit noticeable differences compared to those created by humans. This research survey offers an inclusive exploration of ATS, covering its challenges, types, classifications, approaches, applications, methods, implementations, processing and preprocessing techniques, linguistic analysis, datasets, and evaluation measures, catering to the needs of researchers in the field.

INDEX TERMS Automatic text summarization, text summarization challenges, text summarization methods, text summarization datasets, text summarization evaluation measures.

I. INTRODUCTION

Summarization is the process of compressing a piece of text into a shorter version, lowering the size of the original text while keeping vital informative aspects and content meaning. Because manual text summarizing is a time-consuming and typically arduous activity, automating the work is gaining popularity and thus serves as a major impetus for academic study [1]. It is a technique for condensing large texts such that the summary contains all of the relevant elements from the original content. Text summarization is a difficult problem in the field of natural language processing (NLP). It aims to make reading and searching for information in huge papers easier by creating smaller ones with no loss of significance. Because of the Internet's fast expansion, automatic text summarization (ATS) technologies have become important to address the issue of information content overload. Since people cannot handle large text volumes manually, they attempt to save time and lower costs through the help of automatic analysis tools. Such methods should enable users to make critical decisions by quickly locating the most important information without having to read the entire document [2], [3].

Early experiments in the late 1950's and early 1960's revealed that text summarization by computer was conceivable, but difficult [4]. The methods established at the time were quite crude, depending mostly on surface-level phenomena such as sentence position and word frequency counts, and were geared toward creating extracts (passages taken from the text and repeated verbatim) rather than abstracts

The associate editor coordinating the review of this manuscript and approving it for publication was Claudia Raibulet.

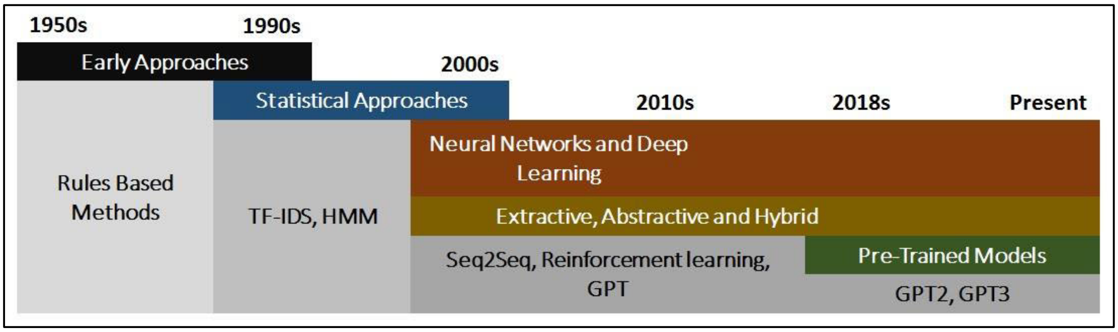


FIGURE 1. Historical timeline of text summarization.

(interpreted portions of the text, newly generated). After a several-decade time away, the expanding existence of vast volumes of online text in corpora and, particularly, on the Web rekindled interest in automated text summarization. During the ensuing decades, advances in NLP, along with significant gains in computer memory and speed, enabled more advanced algorithms to be developed, with highly promising outcomes [5].

With the growth of the internet and big data, people are becoming overwhelmed by the vast amount of information and documents available on the internet. Many researchers are motivated to develop technological approaches that can automatically summarize texts as a result of this. ATS provides summaries that incorporate all essential information from the original material and include crucial sentences [6], [7].

There are several periods in the evolution of text summarization, which show how it has changed through time. Early techniques from the 1950s through the 1990s were rule-based and introduced extractive methods for choosing essential sentences based on factors like sentence length, location, or keywords. In the late 1990s and early 2000s, statistical approaches for text summarization emerged that combined Term Frequency-Inverse Document Frequency (TF-IDF) and statistical methods with machine learning (ML) algorithms. The introduction of neural networks and deep learning (DL) in the 2010s led to a paradigm change by enabling abstractive summarization models like Sequence-to-Sequence (Seq2Seq) and transformers like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), which are frequently improved through reinforcement learning. Hybrid models that included extractive and abstractive components also intended to improve summary coherence. Since 2018, pre-trained language models like GPT-2 and GPT-3 have radically changed text summarization by providing cutting-edge outcomes by fine-tuning on summarization datasets. Figure 1 presents the historical timeline of text summarization since 1990s to present.

Research on ATS plays an important role in today's information-driven world. By offering succinct summaries of lengthy textual content and assisting with speedy understanding and decision-making, ATS systems successfully combat information overload. These solutions increase time effectiveness, accessibility for people with impairments, and content delivery customization. Better search and retrieval, cross-linguistic comprehension, and information extraction are all made possible by ATS, which benefits industries including corporate intelligence and journalism. It also contributes significantly to the development of AI and natural language processing systems, making it a critical field of research for improving our interaction with textual data and information management.

In the present technological era, there is a significant increase in textual data in digital form and it is continuously multiplying. Automatic summarization systems provide convenience to deal with lengthy text data effectively in a time-efficient way. These systems strive to generate summaries that are thorough, succinct, and fluent, while still maintaining all significant information included in a topic. Text summarizing is used in a variety of applications, including search engine snippets created as a consequence of a document search and news websites that generate condensed news in the form of headlines to aid surfing [6], [8].

Figure 2 depicts the entire ATS system, which consists of the following components:

a. The Source Document must be summarized. It might be a single document or a collection of papers.
b. Automatic Text Summarizer This phase is divided into sub-phases that include pre-processing, processing, and post-processing.

This study compiles and synthesizes existing knowledge, research, and information available in the public domain. The procedure and explanation of how data is collected for this study is:

1. **Literature Review and Secondary Data Analysis:** The primary method of data collection for this study is a comprehensive literature review. We gather data from a wide range of sources, including academic journals, conference proceedings, books, reports, and online resources, all of which are publicly available. We do not conduct experiments or interviews but instead rely on the existing body of literature and research on ATS.

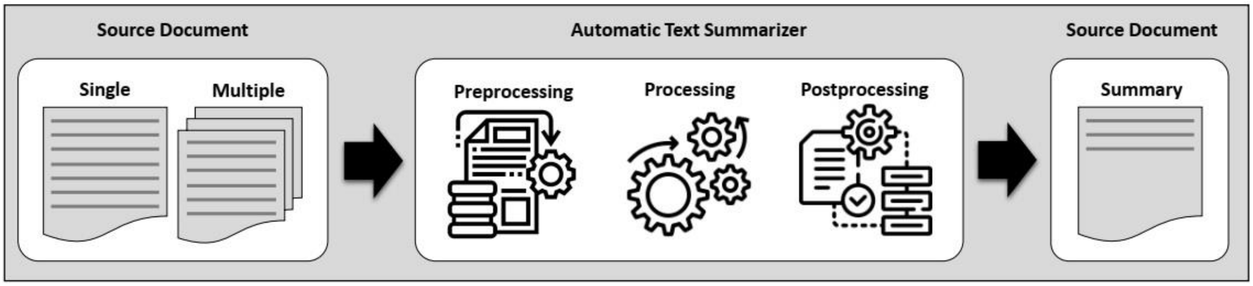


FIGURE 2. The process of automatic text summarization.

2. **Selection and Compilation:** During the literature review process, we select and compile relevant research papers, articles, and resources related to ATS. We gather information on various aspects of ATS, including challenges, types, classifications, approaches, applications, methods, and linguistic analysis, datasets, and evaluation measures.
3. **Synthesis and Summarization:** The collected data, which consists of findings, methodologies, key concepts, and insights from existing literature, is then synthesized and summarized in a structured manner within this study. We use this data to provide a comprehensive overview of the ATS landscape.
4. **Organization and Presentation:** To ensure clarity and accessibility, we organize the collected data into sections, sub-sections, and subheadings. Each section of the paper is dedicated to a specific aspect of ATS, making it easy for readers to navigate and access relevant information.
5. **Citation and Referencing:** Proper citation and referencing of the sources are crucial throughout the study. We acknowledge the original researchers and studies from which we gather data, providing citations to give credit to the original sources.

Overall, the data collection process for this study relies on the systematic gathering, selection, synthesis, and summarization of existing knowledge and research on ATS.

This study primarily focuses on providing an extensive overview of the field of ATS rather than presenting a specific research methodology with distinct stages. The paper serves as a survey and comprehensive reference guide.

The primary purpose of the paper is to:

1. **Introduce the Field:** The study begins by introducing the field of ATS, its components, and its significance.
2. **Provide an In-Depth Exploration:** It explores various aspects of ATS, including challenges, types, classifications, approaches, applications, methods, and linguistic analysis, datasets, and evaluation measures. Each of these sections contributes to a comprehensive understanding of ATS.
3. **Compile Relevant Information:** The study compiles and presents information, categorizing and summarizing research findings, methodologies, and key concepts related to ATS.
4. **Highlight Real-World Implementations:** It also highlights real-world implementations and tools in the field of ATS.
5. **Discuss Evaluation Metrics:** The study discusses evaluation metrics used to assess ATS system performance.

It offers valuable insights, categorization, and an overview of ATS-related topics, serving as a reference for researchers and practitioners in the field.

The remainder of this paper is structured so that Sections II and III present the ATS challenges and Types following each. The ATS classification is covered in Section IV, whereas Section V presents the ATS approaches. The applications and methods of ATS are presented in Sections VI and VII. The implementation of ATS systems is covered in Section VIII. The ATS processes are described in Section IX. Section X presents the linguistic analysis of ATS. Sections XI and XII contain the datasets and evaluation metrics that were used to analyze ATS, respectively.

II. AUTOMATIC TEXT SUMMARIZATION CHALLENGES

ATS systems are rising daily but still, there are many limitations and challenges in ATS. ATS and traditional Text Summarization (TS) face some of the same challenges, such as problems with content selection, coherence, and informativeness. However, we aim to address the wider range of difficulties unique to ATS rather than only highlighting commonalities. When summarizing several documents, dealing with language quirks, and assuring the creation of logical and contextually appropriate summaries, ATS includes complexity that is absent with TS. The study emphasizes the unique character of ATS research and its usefulness in solving real-world summarizing demands by addressing both common issues and those specific to ATS. This method acknowledges the larger context of text summarization while giving readers a well-rounded view of the difficulties encountered in the ATS domain. Some of these are discussed in the subsequent. Figure 3 illustrates the challenges associated with ATS.

A. CONTROLLING THE OUTPUT

The majority of ATS systems deal with textual data as input and output. It is necessary to provide new summarizers in which the input may be meetings, sounds, videos, and so on, and the output can be in a format other than text. For example,

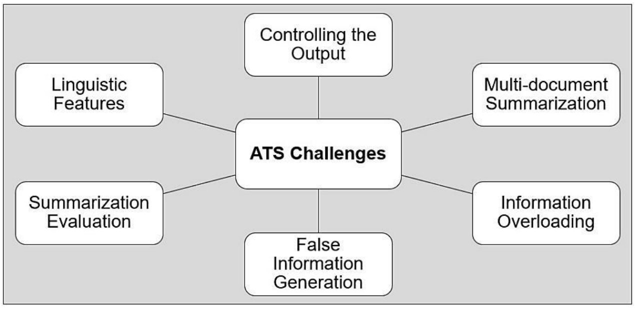


FIGURE 3. Challenges associated with ATS.

the input may be text, and the output could be statistics, tables, visuals, visual rating scales, and so on. ATS systems that allow for the depiction of summaries will assist users in obtaining the necessary material in less time [7], [9].

B. MULTI-DOCUMENT SUMMARIZATION

Multi-document summarizing has greater obstacles than single-document summarization [10]. It addresses challenges such as many documents with duplicate information, multiple document compression, and the speed of sentence selection with its extraction [11]. These challenges are addressed through the use of statistical tools and optimization strategies [12]. Keeping relevance and redundancy under control while summarizing the content is a critical challenge for any ATS system [13].

C. INFORMATION OVERLOADING

One of the most pressing issues today is information overload, which has necessitated the development of increasingly complex and powerful summarizers as a result of the fast growth of the Internet. Recent scientific understanding and more powerful computers have created a new challenge, allowing us to solve the information overload problem, or at the very least postpone it and reduce its harmful impact [14].

D. FALSE INFORMATION GENERATION

In ATS systems, several challenges may be divided into extractive and abstractive strategies. Abstractive approaches, on the other hand, construct an internal semantic representation first and then generate a summary using language processing techniques. A summary like this might include terms that are not in the original report. Existing abstractive text summarization methods, on the other hand, are well-known for generating false information. This might happen at the entity level (additional entities are formed) or at the entity relation level (extra entities are generated) (context in which entities occur is incorrectly generated) [15], [16].

E. SUMMARIZATION EVALUATION

Manual and automatic techniques of summarization evaluation (in intrinsic evaluation, the summary quality is directly based on an examination of the summary, whereas in intrinsic evaluation, the summary quality is determined based on how useful summaries are for a certain job) are used [17]. In the subject of text summarizing research, summary evaluation is a difficult problem to solve. To examine the quality of the ATS systems that created them, the automatically generated summaries must be assessed. The issue with comparing the system summary to an "ideal summary" is that the ideal summary is difficult to define. The human summary might be from the article's author, a judge tasked with creating an abstract, or a judge tasked with extracting sentences. The performance of the ATS system is typically compared to various baseline systems, such as using leading sentences from the input document or using common text summarizers like LexRank [18], TextRank [19], MEAD [20], and so on.

F. LINGUISTIC FEATURES

Linguistic features are mostly employed in the input content to identify relevant sentences and phrases (s). The text summarizing literature employs both word and sentence-level features [21]. Linguistic features are necessary to identify new linguistic and statistical features for sentences and words that can semantically extract essential sentences from the source document(s) [7]. Furthermore, determining the appropriate weights for different attributes is critical since the final summary's quality is dependent on it [22].

III. AUTOMATIC TEXT SUMMARIZATION TYPES

ATS systems are mainly divided into four types that are illustrated in Figure 4 and discussed in the subsequent.

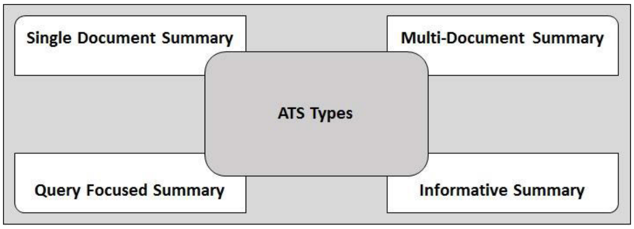


FIGURE 4. Types of ATS system.

A. SINGLE DOCUMENT SUMMARY

Single-document summary reduces a source text to a compacted, shorter version that retains all of the important information or another word, the approach of presenting the primary material of a single document [11].

B. MULTI-DOCUMENT SUMMARY

Multi-Document Summarization is the process of gathering important information and filtering out superfluous information from a series of documents to represent them with a short piece of text. Extractive and abstractive summarizing are two popular techniques for multi-document summarization [12]. Multi-Document is a useful information aggregation tool that creates an interesting and succinct summary from a collection of topic-related publications [23].

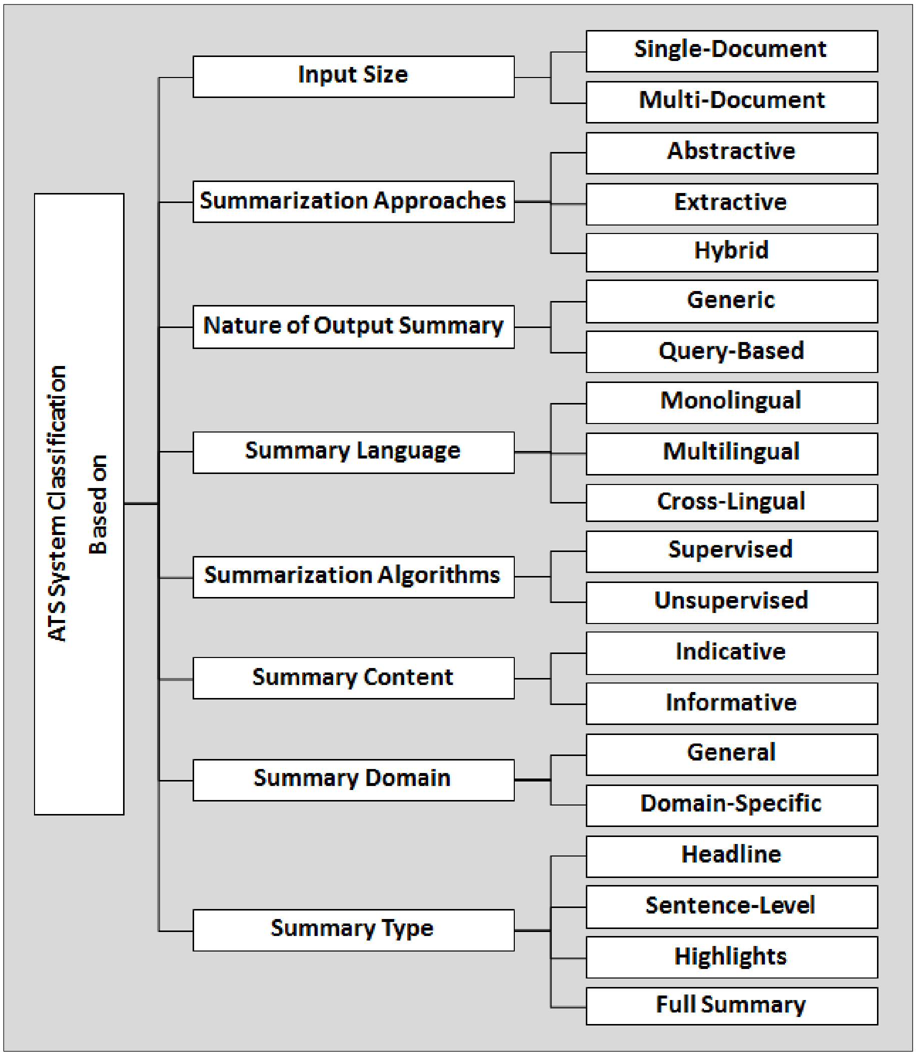


FIGURE 5. ATS system classification.

C. QUERY-FOCUSED SUMMARY

Query-focused summarization models try to extract the most important information from a source text and organize it into a summary that can answer a question. A query-based summary highlights the material most relevant to the original search query, whereas a general summary provides an overview of the document's content [24]. Query-based summaries are also known as query-focused, topic-focused, or user-focused summaries [7].

D. INFORMATIVE SUMMARY

An instructive summary should be neutral, i.e. "presenting the concepts in the original material without personal opinions." To give objective reports of factual information, informative summaries are advised for scientific, and non-fictional publications. An instructive summary [25] comprises all of the relevant information and concepts from the original book, and it covers all of the text's subjects. An instructive summary's goal is to convey the major points of the original material without going into too much detail [26].

IV. AUTOMATIC TEXT SUMMARIZATION CLASSIFICATION

ATS systems are classified basically into eight classes [9] based on different criteria that are presented in Figure 5.

A. CLASSIFICATION BASED ON THE INPUT SIZE

Based on the input size, ATS is classified into two groups: Single-document Summarization (SDS) and Multi-document Summarization (MDS). The number of source documents needed to construct the target summary is referred to as the input size. SDS generates a summary from a single text document, intending to shorten the input content while maintaining the essential information. On the other hand, the summary in MDS is created from a series of input documents, to remove repeated information from the input documents [27]. MDS is more complicated than SDS, including difficulties such as temporal relatedness, redundancy, compression ratio, coverage, and so forth [28].

B. CLASSIFICATION BASED ON THE TEXT SUMMARIZATION APPROACH

There are three types of classification based on text summarization approaches that are Extractive, Abstractive, and Hybrid. The extractive text summarizing method would choose the most essential sentences from the input document(s) and then concatenates them into the output summary. The input document(s) are represented in an intermediate representation in the abstractive text summarizing technique, and the output summary is constructed from this representation. Abstractive summaries, unlike extractive summaries, are made out of sentences that are not the same as the original document's sentences. The extractive and abstractive processes are combined in the hybrid text summarization methodology [29].

C. CLASSIFICATION BASED ON THE NATURE OF THE OUTPUT SUMMARY

The two primary classifications of text summarization depending on the nature of the output summary are generic and query-based. A generic text summarizer gathers key information from one or more input documents to give a broad overview of their contents [7], [30]. A query-based summarizing refers to a multi-document summarizer that works with a set of homogenous documents extracted from a huge corpus as a consequence of a query [30]. The resulting summary then includes items linked to the query. A query-based summary highlights the material most relevant to the original search query, whereas a general summary provides an overview of the document's content [24]. The query-based summary is also known as a query-focused summary, topic-focused summary, or user-focused summary [7].

D. CLASSIFICATION BASED ON THE SUMMARY LANGUAGE

ATS can be classified as Monolingual, Multilingual, or Cross-Lingual based on the summary language. When the source and destination papers are written in the same language, the summarizing system is monolingual. When the source material is written in many languages (e.g., Arabic, English, French, etc.) and the summary is likewise created in these languages, the summarizing system is multilingual. When the source material is written in one language (for example, English) and the summary is created in another (for example, Arabic or French), the summarizing system is cross-lingual [14].

E. CLASSIFICATION BASED ON THE SUMMARIZATION ALGORITHM

Based on summarizing algorithms, ATS may be classed as Supervised or Unsupervised. A training step is required for the supervised algorithm, which requires annotated training data. Because manual annotation of the training data necessitates human work, the latter is difficult to develop and costly. The unsupervised method, on the other hand, does not require a training phase or training data [31].

F. CLASSIFICATION BASED ON THE SUMMARY CONTENT

Based on the summary content, ATS may be characterized as Indicative or Informative. The overall concept or information about the original material is all that is contained in an indicative summary [25]. As a result, it's utilized to figure out what the input text is about (i.e., what themes are discussed) and to notify the user of the source material [32]. An indicative summary's objective is to tell readers about the scope of the input content so that they may determine whether or not to read the full material. An informative summary, on the other hand, provides all of the relevant information and concepts from the original text [25], therefore it covers all of the book's themes [22]. An instructive summary's goal is to convey the major points of the original material without going into too much detail [32].

G. CLASSIFICATION BASED ON THE SUMMARY TYPE

ATS can be classified as Headline, Sentence-Level, Highlights, or Full Summary based on the summary kinds. The length of the produced summaries varies depending on the ATS system's objective. The headline generated via headline creation is frequently less than a phrase. A sentence-level summary takes the input text and creates a single sentence, which is generally an abstractive sentence [33]. A highlights summary is a telegraphic-style, very compressed summary that is often in the form of bullet points. The highlights summary gives the reader a quick rundown of the most important information in the input document(s) [34]. Finally, the desired summary length or compression ratios are frequently used to direct the development of a comprehensive summary.

H. CLASSIFICATION BASED ON THE SUMMARIZATION DOMAIN

ATS can be classed as General or Domain-Specific summarization based on the summarization domain. The generic ATS system, also known as a domain-independent ATS system, summarizes content from several domains. The domain-specific ATS system, on the other hand, is designed to summarize papers from a given domain (e.g. medical documents or legal documents) [9].

V. AUTOMATIC TEXT SUMMARIZATION APPROACHES

Abstractive, extractive, and hybrid text summarization are the three major approaches of the ATS system. These are further split down into subcategories, as seen in Figure 6.

A. ABSTRACTIVE TEXT SUMMARIZATION APPROACHES

Abstractive Text Summarization is the process of creating a brief and succinct summary of a source text that captures the main points. The produced summaries may include

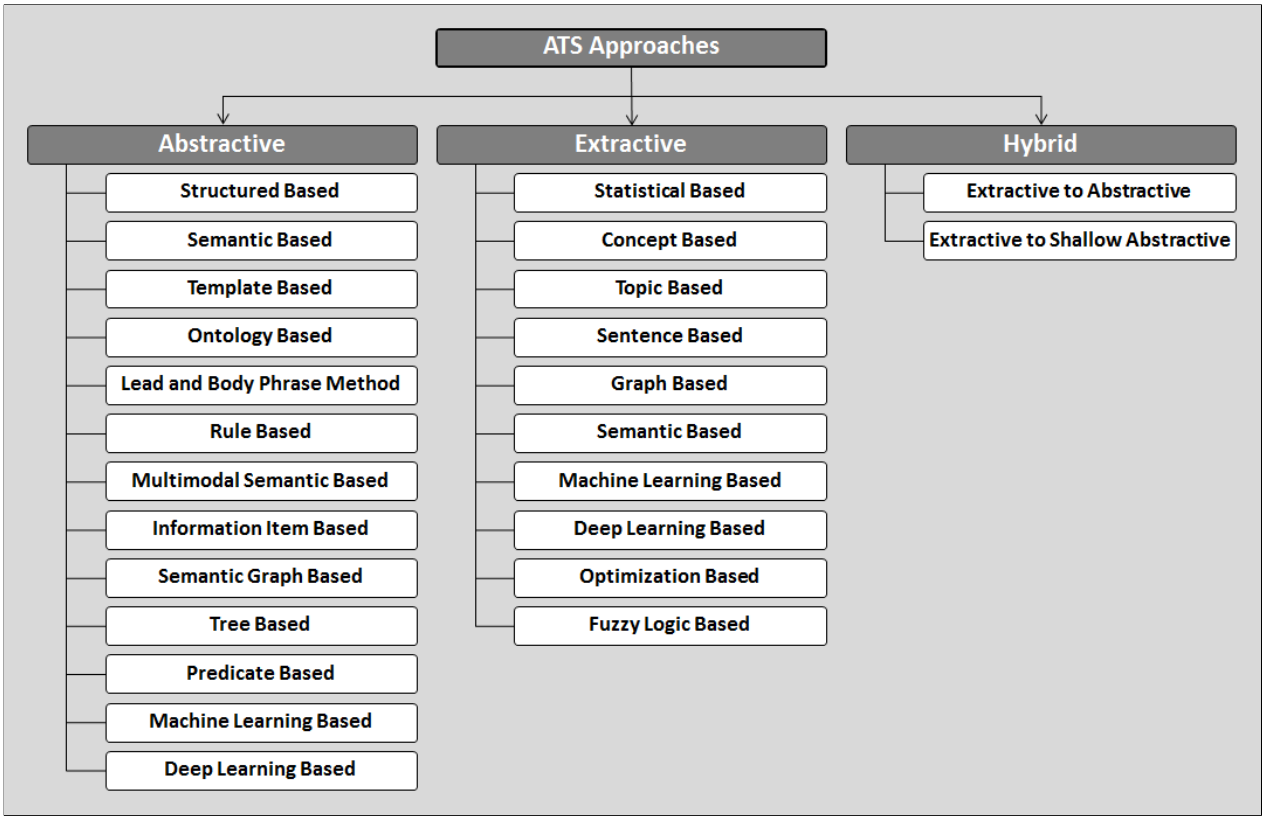


FIGURE 6. ATS approaches.

additional phrases and sentences not found in the original text [35]. In other words, an abstractive text summarization system creates new text containing phrases, sentences, or words that did not occur in the original document while retaining the main meaning of the original. In terms of cohesiveness, readability, and redundancy, Abstractive text summarization attempts to provide high-quality summaries. As a result, this is a difficult assignment since it provides summaries that resemble or approximate those produced by humans [36].

However, as there are merits of these approaches, there are some demerits as well, both are:

Merits:

- **Human-Like Summaries:** Summarization that is abstract resembles human writing.
- **Reduction in Length:** It efficiently cuts down on-the-go content.
- **Paraphrasing and Creativity:** creates a variety of interesting summaries.
- **Handling New Information:** Can provide context outside of the original text.
- **Handling Ambiguity:** Clarifies ambiguities for readability.
- **Content Compression:** Effectively condenses vast amounts of data.
- **Multilingual Support:** Convertible to different languages.
- **Adaptability:** It may be tailored for certain areas.
- **Interpretable Summaries:** Produces clear, user-friendly summary.
- **Improved User Engagement:** Increases user engagement and comprehension.

Demerits:

- **Quality Variability:** Depending on the output quality, summaries may be erroneous or badly written.
- **Complexity:** Compared to extractive techniques, abstractive models are frequently more resource- and complexity-intensive.
- **Training Data Requirements:** To train efficient models, great processing power and large datasets are required.
- **Domain Dependence:** Performance may differ across various subjects and areas.
- **Handling Rare Words:** Has trouble using specialized or uncommon terminology.
- **Content Omissions:** Can leave out crucial data or information.
- **Reference Bias:** Even if inaccurate, it may unintentionally create summaries that resemble training data.
- **Evaluation Challenges:** Absence of standardized measures for abstractive summarization assessment.
- **Ethical Concerns:** Possibility of producing material that is incorrect or biased.

- **Linguistic Challenges:** Difficulty using metaphors, idiomatic idioms, and subtle terminology.

While delivering clear, logical, and useful summaries, abstractive text summarizing methods have obstacles in terms of complexity, possibility for error, and resource needs. The individual application and trade-offs between content preservation and summary quality determine whether to use abstractive or extractive approaches.

1) STRUCTURED-BASED APPROACH

It selects the most important bits of the original documents mostly using DL algorithms. This section might include the document's template, lead and body phrases in the text, extraction rule structure, tree and template-based structure, and so on [37].

2) SEMANTIC-BASED APPROACH

A popular semantic-based extractive ATS technique is latent semantic analysis (LSA). The representation of text semantics by LSA, an unsupervised method, is based on the observed co-occurrence of terms. Every LSA-based extractive summarizer's sentence scoring process begins with the creation of the input matrix (term-to-sentence matrix) and continues with the application of Singular Value Decomposition (SVD) on the input matrix to determine the links between terms and sentences. Other semantic-based ATS methods include explicit semantic analysis and semantic role labeling (SRL) (ESA) [38].

3) TEMPLATE-BASED APPROACH

Human summaries contain common sentence forms that can be referred to as templates for specific areas (such as meeting summaries). By using the information in the input text to fill the slots in the appropriate pre-defined templates, the abstractive summary may be generated based on the genre of the input text. The text samples that fill the template slots are selected using extraction rules and linguistic patterns [39].

4) ONTOLOGY-BASED APPROACH

Each domain has its information structure that may be represented by a knowledge dictionary similar to ontology. Many articles are associated with certain domains. The fundamental concept is to extract the necessary data from the input text to construct an abstractive summary using an ontology [40].

5) LEAD AND BODY PHRASE METHOD APPROACH

This method is based on the "insert and replace" process, which uses core sentences to replace the leading phrase and comparable syntactic head chunks at the beginning. Structures for the lead, body, and supplements are used to represent text. Material is chosen depending on how much text there is that has the same lead and body format. Use of the insertion and replacement approach is used to generate summaries. As it deals with the semantics of the sentences, this technique has the disadvantage of grammatical errors while producing summaries, which is advantageous for semantic-based summaries [37].

6) RULE-BASED APPROACH

Rule-based text summarization uses a set of predetermined rules or algorithms to summarize lengthy texts. This method aims to highlight the most crucial details from the original text and deliver them concisely. When it's necessary to swiftly extract crucial information from massive amounts of text, rule-based text summarization might be helpful. Nevertheless, because it uses predetermined criteria rather than natural language comprehension, it might not always generate the most accurate or thorough summary [39].

7) MULTIMODAL SEMANTIC-BASED APPROACH

To represent the topic (images and text data) of one or more documents, a semantic unit is produced that extracts the subject matter and correlation among the topics. The key topic is scored based on various criteria, and the selected topics are then created as sentences to construct a summary. The disadvantage of this approach is that the resulting summary is physically validated by people, which might have been done automatically [41].

8) INFORMATION ITEM-BASED APPROACH

Using the original text's sentences as a starting point, this approach creates the information for the summary from an abstract representation of the original text. The tiny instinct of the associated information in the text is an abstract representation of the data object. Information item retrieval, sentence production, sentence selection, and summary creation are the three components of the framework that makes up the approach. Syntactic analysis of the input data is carried out by a parser during the information item retrieval phase, and the verb's topic and object are chosen. As a result, the positioned topic-verb-object triple is used to define the information item-based method. Using the average document frequency value and the information item retrieval, a sentence is generated in the sentence creation phase. Last but not least, a summary generation step yields precise, pertinent information while maintaining the original meaning of the source content [37].

9) SEMANTIC GRAPH-BASED APPROACH

This method generates a summary by generating a semantic network on the source material termed a Rich Semantic network (RSG), compressing the semantic graph, and then extracting the full abstractive summary from the condensed semantic graph. There are three steps in this process. The input document makes use of RSG conceptually. RSG depicts the nodes of the graph, which are the words and verbs from the input document, and the edges connecting them as their topological and semantic relationships. The second section uses heuristic rules to compress the constructed semantic network of the input material. The abstractive summary is produced in the third section using the compressed RSG. This section gets the RSG that has been semantically expressed and generates the summary [42].

10) TREE-BASED APPROACH

This approach uses a dependency tree to describe the text and information from the source text. Applying a variety of algorithms results in the absorption of the information for summary creation. This method's flaw is that it lacks a perfect model with an accurate description of information retrieval [43].

11) PREDICATE-BASED APPROACH

Predicate-based text summarizing is an NLP approach that seeks to construct a summary of a given text by finding and extracting the text's most essential predicates or propositions. Predicates capture the key activities or occurrences reported in a sentence, and their meaning can be expressed by the sentence's predicates [44].

12) MACHINE LEARNING BASED

ML has transformed the area of text summarizing by allowing for the automatic, accurate, and rapid summation of enormous amounts of text. A ML model is trained on a huge corpus of text in this technique to detect significant words, crucial sentences, and essential information in the text. This method has also been used successfully to summarize research publications [45].

13) DEEP LEARNING BASED

Due to its capacity to recognize intricate patterns and relationships in the data, DL has demonstrated considerable promise in text summarization. Encoder-decoder models, such as the sequence-to-sequence (Seq2Seq) model, are a well-liked method that has been effectively used for text summarization [46], [47].

B. EXTRACTIVE TEXT SUMMARIZATION APPROACHES

The process of extractive text summarizing involves choosing and extracting significant sentences or phrases that represent the substance of the original text to automatically create a summary of a text document. This method involves picking a selection of sentences from the original text and arranging them in a logical sequence to make the summary. This contrasts with abstract text summary, which entails coming up with new phrases to accurately convey the original text's meaning. In many different applications, such as document summarizing, chatbot answer creation, and news summary, extractive text summarization is often employed. Extractive text summarization aims to provide a summary of the text that preserves the important details and is simple to comprehend [48], [49]. While content preservation and simplicity are two benefits of extractive text summarization, it also has drawbacks relating to coherence and repetition. Depending on individual use cases and desired summary qualities, one must choose between extractive and abstractive approaches.

Merits:

- **Preservation of Source Content:** Direct selection and inclusion of phrases from the original text using extractive methods guarantees the retention of important details in the summary.
- **Reduced Risk of Information Loss:** Since extractive summarization uses content already in existence, there is less chance of leaving out crucial information or including errors.
- **Simplicity:** Extractive techniques are computationally efficient because they are frequently easier to execute than abstractive approaches.
- **Interpretability:** Since phrases from the original text are used to create summaries, they are easier to grasp and comprehend.
- **Fewer Training Data Requirements:** In comparison to abstractive models, extractive models could need fewer training datasets.

Demerits:

- **Redundancy:** Repetitive summaries may result from extracted phrases containing redundant information.
- **Lack of Coherence:** It's possible that sentences that were extracted did not make sense, leading to summaries that lacked general coherence and organization.
- **Inability to Paraphrase:** The inability of extractive approaches to rephrase or rewrite phrases limits their capacity to provide succinct, comprehensible summaries.
- **Limited to Existing Content:** Extrapolative models are unable to offer knowledge or insights that go beyond what is included in the original text.
- **Coverage Issues:** If the source material is extensive or if crucial information is dispersed among several phrases, they can miss important elements.
- **Sensitivity to Input Order:** The summary may be impacted by the sequence of the phrases in the original text, sometimes producing different results.
- **Difficulty with Pronouns:** Extraction models could have trouble correctly resolving pronoun references.
- **Lack of Abstraction:** In situations when abstraction is necessary, extractive approaches might not be able to produce succinct summaries.

1) STATISTICAL BASED

These techniques use statistical analysis of a collection of attributes to extract significant phrases and words from the source text. According to Gupta and Lehal [22], the "most important" statement is the one that is "most favorably positioned," "most frequent," etc. A statistical-based extractive summarizer's sentence scoring steps are as follows [50]: a) selecting and computing some statistical and/or linguistic features, then giving them weights and, b) giving each sentence in the document a final score that is determined using a feature-weight equation i.e. all the selected features' scores are computed and summed.

2) CONCEPT BASED

These techniques use external knowledge bases such as WordNet, HowNet, Wikipedia, etc. to extract concepts from a text. Then, rather than using words, the importance of sentences is determined using the concepts retrieved from the external knowledge base HowNet. According to Moratanch and Chitrakala [21], a concept-based extractive summarizer's sentence scoring steps entail a) retrieving concepts from an external knowledge base, b) creating a conceptual vector or graph model to illustrate the relationship between concepts and sentences, and c) using a ranking algorithm to score the sentences.

3) TOPIC BASED

These techniques focus on determining the theme, or the primary subject (i.e., what the document is about), of a text. Term Frequency, TF-IDF, lexical chains, and topic word approaches, in which the topic representation consists of a straightforward table and their related weights [51], are some of the most popular techniques for topic representations. The steps in the processing of a topic-based extractive summarizer are: a) converting the input text into an intermediate representation that captures the topics discussed in the input text; and b) giving each sentence in the input documents an importance score following this representation.

4) SENTENCE BASED

Sentence-based text summarizing is a method for condensing a larger text into a shorter summary by highlighting the key phrases that best express the content's primary concepts. To find important phrases and extract the most pertinent data, this technique uses ML algorithms and natural language processing. Readers who are pressed for time or who need to rapidly understand the major ideas of a longer work will find the resultant summary to be easier to understand and absorb [52], [53].

5) GRAPH BASED

This approach is used to construct a summary of a larger piece of text by describing the text as a graph and then extracting relevant phrases based on the graph structure. Nodes in the graph represent sentences in this technique, while edges reflect correlations between them, such as co-occurrence or similarity. The computer then evaluates each sentence's position in the network and its links to other phrases to determine its significance score. Then, the key phrases are picked out to create a summary that captures the essence of the original text [54]. This method is frequently applied when a text is too lengthy for human summation or when a more impartial summary is necessary. When there is a lot of repetition in the original text or when the same concepts appear in several phrases, graph-based summarization can also be helpful [9], [38].

6) SEMANTIC BASED

Semantic-based text summarizing is a method for creating a summary of a larger text by examining the words and phrases that are used in the text as well as their context. To find the most crucial ideas and topics in the text and to extract the most pertinent data, this method uses ML and natural language processing techniques [55]. The standard method for semantic-based summarization includes text preparation, entity recognition, and concept extraction. The text is cleaned up during the text preparation stage so that it is ready for analysis. The algorithm recognizes and classifies named entities like persons, locations, and organizations during the entity recognition stage. Based on the connections between the recognized entities and the words and phrases used in the text, the algorithm extracts essential concepts and themes from the text in the concept extraction stage [56].

7) MACHINE LEARNING BASED

At the sentence level, these strategies transform the summarization problem into a supervised classification challenge. Using a training set of documents (i.e. a collection of documents and their associated human-generated summaries), the system learns by example to categorize each sentence of the test document as a "summary" or "non-summary" class. The sentence scoring steps for the machine-learning-based summarizer are as follows [21]: a) extracting features from the preprocessed document (based on multiple features of sentences and words), and b) feeding the extracted features to a neural network, which produces a single value as an output score.

8) DEEP LEARNING BASED

Kobayashi et al. [57] present a summarization technique based on embedding that uses document-level similarity. A word's embedding represents its meaning. A document is thought of as a bag of sentences, and a sentence is thought of as a bag of words. The challenge is formalized as maximizing a sub-modular function defined by the negative sum of the closest neighbor's distances on embedding distributions. According to Kobayashi et al., document-level similarity can determine more complicated interpretations than sentence-level similarity. The author in [58] offers an ATS system for single document summarization based on a reinforcement learning algorithm and encoder-extractor network architecture's Recurrent Neural Network (RNN) sequence model. A sentence-level selective encoding approach is used to choose the key characteristics, and then summary sentences are retrieved.

9) OPTIMIZATION BASED

Optimization-based text summarizing is a strategy for creating a summary of a lengthier piece of text by presenting the task as an optimization issue. The purpose is to choose a group of phrases from the original text that maximizes a specified objective function, such as covering key information while minimizing duplication [59]. In this method, candidate sentences that could be included in the summary are initially found by the algorithm. The work of summarizing is then formulated as an optimization problem, with the objective function stated in terms of the desirable characteristics of the summary, such as its length, coherence, and in formativeness. The best collections of words that fulfill the objective function are then obtained by applying the algorithm's solution to the optimization issue.

10) FUZZY LOGIC BASED

Using fuzzy logic concepts to assess the relevance of phrases to the primary theme of the text, fuzzy logic-based text summarizing is a technique for creating a summary of a lengthy piece of text. This method evaluates each phrase according to how closely it relates to the primary topic, which is represented as a fuzzy set [60]. The algorithm uses fuzzy logic operations, such as fuzzy intersection, fuzzy union, and fuzzy complement, to calculate the degree of relevance of each sentence. The sentences with the highest degree of relevance are then selected to create the summary. Fuzzy logic-based summarization can be especially effective when the primary theme of the text is unclear or the text contains confusing or imprecise terminology. It is especially beneficial when the relevance of individual phrases fluctuates based on the context or the reader's choices [53].

C. HYBRID TEXT SUMMARIZATION APPROACHES

A method for creating text summaries that include several different approaches is known as hybrid text summarizing. Using the advantages of various techniques, hybrid text summarization seeks to create high-quality summaries that highlight the key points of a text. Extractive and abstractive text summarizations are the two main types. In contrast to abstractive summarization, which creates new text that accurately captures the meaning of the original text, extractive summarization involves choosing the most crucial sentences or phrases from a text and combining them into a summary [61]. Techniques from both extractive and abstractive summarizing are frequently used in hybrid text summarization methodologies. To guarantee that it includes all of the most crucial information, a hybrid method, for instance, can use a neural network to create an initial summary that is subsequently improved upon using extractive summarization techniques. One benefit of hybrid text summarizing is that it makes it possible to provide a more accurate and nuanced summary of a text than would be possible with only one approach. However, compared to using a single method, it can also be more computationally expensive and complex [9], [25].

The advantage of hybrid text summary techniques is that they combine the best aspects of extractive and abstractive methods, producing summaries that are more logical and insightful. They do, however, provide difficulties due to complexity, resource constraints, and customization requirements. The specific summarizing goal and the trade-offs between content preservation and summary quality determine which hybrid technique is best. Here are some of the merits and demerits of hybrid text summarization approaches.

Merits:

1. **Combining Strengths:** Hybrid approaches combine extractive and abstractive techniques, using the benefits of both to provide summaries that are more accurate and comprehensive.
2. **Improved Coherence:** By using abstractive approaches to rearrange and organize retrieved phrases, they can provide summaries with improved coherence.
3. **Content Preservation:** Hybrid approaches frequently do exceptionally well at maintaining key source information while supplying abstractive aspects for clarity and conciseness.
4. **Reduced Abstraction Risk:** They reduce the chance of include errors or misunderstandings in summaries by combining extractive and abstractive approaches.
5. **Versatility:** A wide range of summarization tasks can benefit from hybrid techniques since they can be tailored to different text kinds and topics.

Demerits:

1. **Complexity:** Because both extractive and abstractive components must be integrated and adjusted, creating hybrid models may be difficult.
2. **Resource Intensive:** For both extractive and abstract parts, they frequently need significant computational resources and training data.
3. **Customization Needs:** It may be essential to fine-tune hybrid models for certain domains or activities, which might lengthen the development process and require more work.
4. **Quality Variability:** The efficacy of the integrated components and the selected parameters can have an impact on the hybrid summary' quality.
5. **Trade-offs:** Finding the correct balance between content retention and abstraction can be challenging, and it relies on the particular requirements for summarizing.

1) EXTRACTIVE TO ABSTRACTIVE

These techniques start by applying an extractive ATS technique, and then they use an abstractive text summarization technique to the extracted phrases. Wang et al. [54] propose a hybrid system called ''EA-LTS'' for lengthy text summarization. The system is divided into two phases: a) the extraction phase, which builds an RNN-based encoder-decoder and employs a pointer and attention mechanism to extract the essential sentences, and b) the abstraction phase, which builds an encoder-decoder and generates summaries.

2) EXTRACTIVE TO SHALLOW ABSTRACTIVE

Text summarizing approaches that seek to construct summaries of a text document includes extractive summarization and shallow abstractive summarization. To construct a summary, extractive summarizing selects and concatenates the most essential sentences or phrases from a document, whereas shallow abstractive summarization modifies the selected sentences to make them more compact and cohesive. These methods begin by employing one of the extractive ATS techniques, and then they employ a shallow abstractive text summarization technique that employs one or more of the information compression techniques, information fusion techniques [62], synonym replacement techniques [63], etc. to the extracted sentences.

VI. AUTOMATIC TEXT SUMMARIZATION APPLICATIONS

ATS is used in many different fields and applications. Applications of ATS are tools that may produce summaries of large documents, publications, or web pages. The most crucial information is extracted by these programs using algorithms, and it is then presented in a condensed manner. Some of the key applications of ATS are shown in Figure 7.

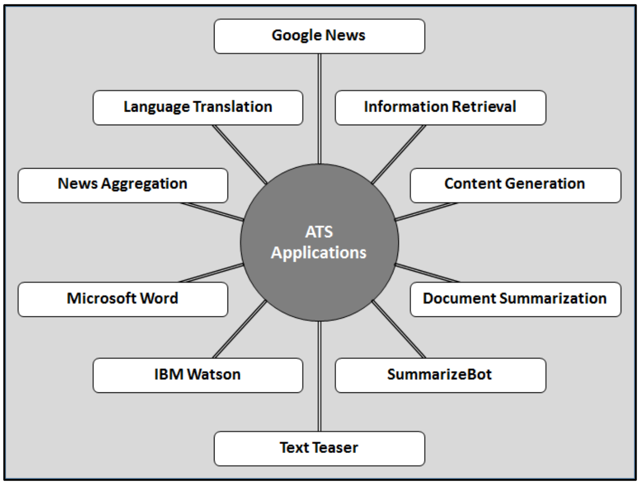


FIGURE 7. Some of the ATS applications.

A. GOOGLE NEWS

Popular news aggregator Google News offers automated text summaries. A summary of the news story is frequently displayed at the top of the page when you click on it in Google News. Automatic algorithms that examine the article's content produced this summary.

B. INFORMATION RETRIEVAL

Search engines may summarize the search results using automatic text summarizing. Users may then choose which results to click on after rapidly understanding the content of each result.

C. CONTENT GENERATION

Automatic text summarizing may be used to create summaries of user-generated information, including product evaluations, social media posts, and consumer feedback. Businesses may use this to understand client sentiment fast and decide what to do.

D. DOCUMENT SUMMARIZATION

Summaries of lengthy documents, such as reports, research papers, and legal documents, may be created automatically by analyzing the language of the document. Those who need to read through a lot of content can save time and effort by doing this.

E. SUMMARIZEBOT

SummarizeBot is an AI-powered text summarizing tool that can quickly summarize any kind of text. It analyses text using natural language processing (NLP) technology to extract the most crucial information. SummarizeBot offers summaries in several formats and supports many languages.

F. TEXT TEASER

Text Teaser is an open-source text summarizing application that creates summaries using ML methods. It can summarize plain text documents as well as web pages and supports many different languages.

G. IBM WATSON

NLP and ATS are only two of the AI-powered capabilities offered by IBM Watson, a cognitive computing platform. Research papers, news stories, and legal documents are just a few examples of the kind of texts that Watson may be used to summarize.

H. MICROSOFT WORD

A built-in tool in Microsoft Word called "AutoSummarize" may automatically create a summary of a document. To provide a summary that is a specific proportion of the original length, this feature employs algorithms to find the text's important ideas.

I. NEWS AGGREGATION

To give users a summary of news stories, automatic text summarization can be utilized in news aggregation applications and websites. Users will be able to rapidly comprehend the story's essential elements without having to read the complete post.

J. LANGUAGE TRANSLATION

Machine translation systems can utilize automatic text summaries to summarize a text's meaning in one language before translating it into another. This might enhance the speed and accuracy of machine translation.

Overall, ATS software may be quite helpful for anyone who needs to swiftly and effectively handle vast quantities of text. It might be a useful tool in a variety of situations where it is necessary to swiftly comprehend huge amounts of text.

VII. AUTOMATIC TEXT SUMMARIZATION METHODS

ATS is a technique for automatically condensing a written document while retaining the most important information. There are different methods of summarizing material mechanically, including:

A. EXTRACTIVE SUMMARIZATION

To construct a summary using this technique, the most pertinent sentences or phrases from the original text are chosen and extracted. The importance and relevance of the selected sentences to the text's primary subjects or themes are often taken into consideration while selecting them [38], [48].

Advantages:

- Preserves the text's original phrasing and grammar
- Maybe more trustworthy than abstractive summarizing since it depends on material already present in the text
- Typically performs quicker and with fewer CPU resources than abstractive summarization techniques

Disadvantages:

- May not capture the essence of the text if key information is dispersed throughout the document and absent from the extracted sentences
- May be limited in its capacity to produce novel insights or new perspectives on the text.
- This may result in awkward or disjointed sentences if extracted sentences are improperly connected to or contextualized.

B. ABSTRACTIVE SUMMARIZATION

The most pertinent sentences or phrases from the original text are chosen and extracted using this technique to produce a summary. The significance and relevance of the selected sentences to the primary issues or themes of the text are typically taken into account while selecting them [40], [49].

Advantages:

- Can produce summaries that convey the text's main ideas even when they aren't stated explicitly in the original work
- Can produce more fluid, cogent sentences than extractive summarization
- Has the potential to produce fresh interpretations of the text that aren't included in the original work.

Disadvantages:

- It requires a more involved natural language generation process
- It may be more error-prone, especially when the text uses ambiguous or complex language
- It may also be slower and use more computational resources.

C. HYBRID SUMMARIZATION

To create summaries, this technique combines the extractive and abstractive processes. Using abstractive techniques to create new sentences that summarize the material that was originally taken from the text is a common way for hybrid summarization [9], [25].

Advantages:

- Combines the benefits of extractive and abstractive summarizing techniques and can result in more thorough and informative summaries than utilizing just one method.
- Since it can start with the most significant sentences in the original text, it can be more effective than pure abstract summarizing.

Disadvantages:

- May be more difficult to implement and need more computing power than only extractive or abstract approaches
- May not always result in summaries that are superior to those created just by extractive or abstractive approaches.

D. NEURAL NETWORK-BASED SUMMARIZATION

With the use of massive text document datasets, D models are trained in this manner to produce summaries. Recurrent neural networks (RNNs) or transformer-based models like BERT or GPT are frequently used in neural network-based summarization, which can be either extractive or abstractive [64].

Advantages:

- They are capable of producing summaries that are extremely accurate and fluent, especially when trained on big-text datasets
- They are efficient and effective in summarizing vast amounts of text
- They may be tailored to particular domains or themes by training on specialized datasets.

Disadvantages:

- Neural network-based summarization requires a substantial amount of training data
- It might be challenging to understand how the model creates summaries.
- Neural networks may overfit training data.

E. LATENT SEMANTIC ANALYSIS (LSA)

This approach makes use of mathematical methods to find word correlations and patterns in the text. The most significant points in the text may be found using LSA, and a summary can be created using those points [65].

Advantages:

- Can be used to create summaries that concentrate on a document's most significant subjects
- Can be useful in summarizing technical or specialized terminology.
- Can uncover patterns and links between words in a text that are not immediately evident.

Disadvantages:

- May not always capture the subtleties of language and meaning in the text.
- Needs a lot of data to train the model and identify the key subjects in the text.

F. GRAPH-BASED SUMMARIZATION

With this approach, the text is visualized as a graph with sentences or phrases acting as nodes and relationships between them as their edges. The graph's most significant nodes are then chosen to provide a summary [66].

Advantages:

- Can be useful in situations where the text contains multiple themes or subtopics that need to be summarized separately.
- Can be more effective than extractive or abstractive summarization methods in some cases.
- Can be effective in identifying the most important content in a text document by visualizing relationships between sentences or concepts.

Disadvantages:

- A well-defined model of the links between phrases or concepts is necessary, which can be challenging to establish in complicated texts.
- Might not always result in legible or cohesive summaries.

The individual use case and the demands placed on the generated summary, such as its length, readability, and correctness, ultimately determine the approach to be used.

VIII. IMPLEMENTATIONS OF AUTOMATIC TEXT SUMMARIZATION SYSTEMS

The value of ATS is found in its capacity to speed up and improve productivity for users who must handle enormous amounts of text. Users may rapidly comprehend the main ideas without having to read the full content by presenting a shortened version of it. For those who have trouble reading or other impairments, ATS can also assist to increase accessibility, making it simpler for them to have access to crucial information. There are various implementations of ATS, some of which are listed in Table 1.

These tools may be applied to a wide range of tasks, including summarizing news stories, academic papers, or legal documents.

TABLE 1. Automatic text summarization implementations.

Implementations	Descriptions
GPT	An OpenAI language model that is capable of extractive and abstractive summarization [67].
TextRank	A graph-based ranking system that can identify a text's most crucial phrases [68].
Summarizer	A Python module that extracts summaries from text using TextRank [69].
BART	A Facebook AI-developed transformer-based model that has already been trained and is capable of abstractive summarization [70].
T5	A Google AI-developed pre-trained transformer-based model that is capable of both extractive and abstractive summarization [71].
NLTK	A Python package that offers several tools for extracting summarization and other aspects of natural language processing [72].
SpaCy	It is a Python package for extractive summarization and natural language processing [73].
Microsoft Cognitive Services	An API for text summarizing is part of the Microsoft Cognitive Services toolkit of APIs and technologies [74].
Hugging Face	It is a Python package that gives users access to several already-trained NLP models, including summarization models [75].

IX. AUTOMATIC TEXT SUMMARIZATION PROCESSING AND PREPROCESSING TECHNIQUES

ATS uses several processing and preprocessing techniques to extract the most significant information from the incoming text and provide a summary [9], [12], [48]. Figure 8 shows some of the ATS processing and preprocessing techniques.

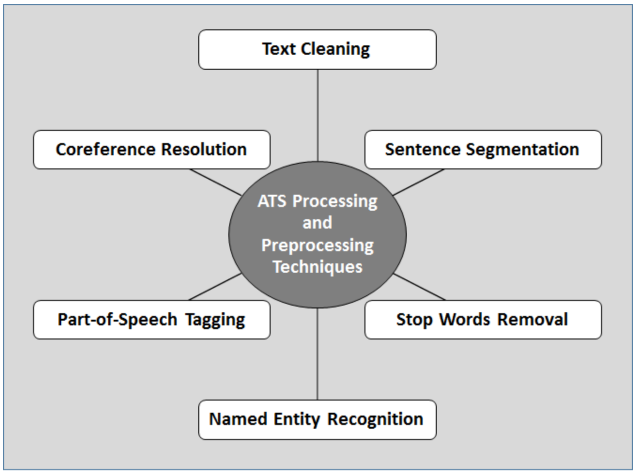


FIGURE 8. ATS processing and preprocessing techniques.

A. TEXT CLEANING

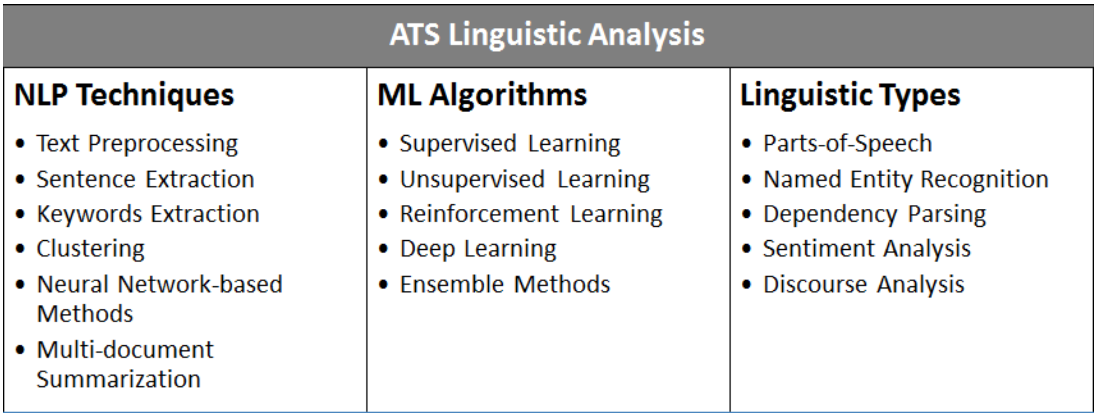
In this preparation method, extra characters, symbols, and punctuation are removed from the text. This lessens noise and raises the precision of later processing methods [76].

B. SENTENCE SEGMENTATION

Using this method, the input text is divided up into separate sentences. This stage is crucial to extractive summarizing since it makes it possible to choose the key phrases for the summary [77].

C. STOP WORDS REMOVAL

Stop words are often used words like "the," "and," "a," etc. that have little to no relevance in the text. Removing these terms helps to clean up the summary and eliminate noise [78].



ATS Linguistic Analysis

NLP Techniques	ML Algorithms	Linguistic Types
• Text Preprocessing	• Supervised Learning	• Parts-of-Speech
• Sentence Extraction	• Unsupervised Learning	• Named Entity Recognition
• Keywords Extraction	• Reinforcement Learning	• Dependency Parsing
• Clustering	• Deep Learning	• Sentiment Analysis
• Neural Network-based Methods	• Ensemble Methods	• Discourse Analysis
• Multi-document Summarization		

FIGURE 9. Automatic text summarization linguistic analysis.

D. STEMMING AND LEMMATIZATION

To eliminate repetition and boost the effectiveness of the summarization process, these approaches include breaking down words into their root forms. Lemmatization is mapping words to their basic form, whereas stemming entails eliminating prefixes and suffixes from words [79].

E. NAMED ENTITY RECOGNITION (NER)

This method involves locating and extracting significant entities from the text, including people, companies, and locations. Summaries that provide more information may be produced using these items [80].

F. PART-OF-SPEECH (POS) TAGGING

This method identifies the parts of speech that each word in a phrase belongs to, such as nouns, verbs, adjectives, etc. The text's significant phrases and sentences may be located and extracted using this information [81].

G. SENTIMENT ANALYSIS

This method involves examining the text's emotional tone. It may be used to determine which sentences in the book are the most uplifting or depressing then incorporate them in the summary appropriately [82].

H. COREFERENCE RESOLUTION

Using this method, all instances of a particular entity in the text are found and replaced with a single reference. This lessens repetition and strengthens the coherence of the summary [83].

These processing and preprocessing methods may be used with various summary strategies, including extractive, abstractive, and hybrid summarization, to produce summaries that highlight the most crucial details from the original text.

X. AUTOMATIC TEXT SUMMARIZATION LINGUISTIC ANALYSIS

ATS includes reducing a larger text to a shorter one while retaining the key points. By identifying the important concepts, entities, and connections inside the text, linguistic analysis plays a critical part in this process [7], [9]. The several methods and varieties of ATS linguistic analysis are depicted in Figure 9.

A. NATURAL LANGUAGE PROCESSING TECHNIQUES

NLP methods are a popular method for linguistic analysis in text summarization. By dissecting human language into its component pieces, such as words, phrases, and sentences, NLP enables computers to analyze and comprehend human language [84]. For automated text summarization, several NLP approaches are utilized. Here are some of the most often utilized methods:

1) TEXT PREPROCESSING

Text preprocessing includes cleaning and getting the input text ready for additional analysis. It entails activities like getting rid of stop words, changing the text's case to lowercase, stemming or lemmatization, and getting rid of punctuation and special characters [85].

2) SENTENCE EXTRACTION

To include the most significant sentences in the summary, the most significant sentences from the input text must be found. Based on several variables, including sentence length, word frequency, and sentence location, this may be done [86].

3) KEYWORD EXTRACTION

This process includes locating the essential phrases or ideas in the supplied text. Graph-based approaches or techniques like TF-IDF can be used for this [87].

4) CLUSTERING

Clustering is the act of assembling phrases or ideas that are related. As a result, a summary that encompasses the text's many facets may be produced [88].

5) NEURAL NETWORK-BASED METHODS

They include training DL models, such as recurrent neural networks (RNNs) or transformers, using massive datasets of summaries created by humans. To create summaries for fresh input texts, these models may then be applied [89].

6) MULTI-DOCUMENT SUMMARIZATION

This process is putting together a summary of several articles or papers that are pertinent to the same subject. Sentence fusion and topic modeling are two methods that may be used to accomplish this [90].

B. MACHINE LEARNING ALGORITHMS

Another method of linguistic analysis in text summarization is to use ML algorithms to discover patterns in the text. These algorithms may be trained on vast datasets of human-generated summaries to understand which sentences or phrases are most useful for summarizing a certain sort of material. For automated text summarization, numerous ML approaches may be utilized [91]. Here are some strategies that are regularly used:

1) SUPERVISED LEARNING

This entails training a ML model with a labeled dataset of input texts and summaries. After that, the model may be used to create summaries for fresh input texts. Decision trees, support vector machines (SVM), and neural networks are common supervised learning techniques used for summarization [92].

2) UNSUPERVISED LEARNING

This includes grouping related words or concepts and generating a summary using clustering or topic modeling methods. For summarization, common unsupervised learning methods include k-means clustering, latent semantic analysis (LSA), and latent Dirichlet allocation (LDA) [93].

3) REINFORCEMENT LEARNING

This is the process of teaching a ML model how to create summaries depending on input from a reward function. The model has been trained to maximize the reward, which is determined by how closely the produced summary matches the reference summary [94].

4) DEEP LEARNING

This entails training deep neural networks on massive datasets of input texts and summaries, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). After that, these models may be utilized to produce summaries for new input texts [95].

5) ENSEMBLE METHODS

These combine numerous ML models to produce a more robust and accurate summary. This may be accomplished through the use of strategies such as bagging, boosting, and stacking.

C. LINGUISTIC TYPES

ATS uses a variety of linguistic types to analyze incoming text and generate a summary [9]. Here are some of the most prevalent linguistic types utilized in automatic text summarization:

1) PART-OF-SPEECH TAGGING

POS tagging entails determining the part of speech (e.g., noun, verb, adjective) of each word in the input text. This data is used in summarization to find the most essential words and verbs that express the text's primary concepts [96].

2) NAMED ENTITY RECOGNITION

NER entails recognizing and categorizing named entities referenced in the text, such as persons, organizations, locations, and dates. This information may be used in summarizing to determine the most significant elements in the text [97].

3) DEPENDENCY PARSING

This is the process of analyzing the links between words in a phrase and determining the principal subject, predicate, and object. This information can be used to identify the major acts and events discussed in the text while summarizing [98].

4) SENTIMENT ANALYSIS

Sentiment analysis entails determining the text's emotional tone. This information may be used to determine the most positive and negative parts of the text during summarization [99].

5) DISCOURSE ANALYSIS

Discourse analysis entails examining the text's structure and organization, such as identifying primary arguments, transitions, and logical links between words. This information may be utilized to determine the most essential ideas and arguments given in the text during summarizing [100].

XI. AUTOMATIC TEXT SUMMARIZATION DATASETS

This section gives an overview of the fundamental materials used to evaluate and compare ATS systems. Among these resources are well-known and standard datasets. As shown in Table 2, this survey covers the most often used benchmarking datasets for ATS system evaluation. It gives a systematic overview of common datasets used for text summarizing software. It has columns for "Dataset" names, which show dataset titles; "No. of Documents," which indicates how many text samples are included in each dataset; "Language," which indicates the language in which the documents are written; "Domain," which describes the subject or topic of the documents; "Single/Multi-Document," which distinguishes between datasets for single-document or multi-document summarization tasks; and "URL," which provides references or links to access the data. Based on their needs for language, domain, and summarizing task-specific datasets, researchers and practitioners may make informed decisions.

TABLE 2. Standard datasets for text summarization.

Dataset	No. of Documents	Language	Domain	Single/Multi-Document	URL
TAC 2008 [101]	48x20	English	News	Multi	https://tac.nist.gov/data/index.html
TAC 2009 [102]	44x20	English	News	Multi	
TAC 2010 [103]	46x20	English	News	Multi	
TAC 2011 [104]	44x20	English	News	Multi	
DUC 2001 [105]	60x10	English	News	Both	https://www-nlpir.nist.gov/projects/duc/data.html
DUC 2002 [106]	60x10	English	News	Both	
DUC 2003 [107]	60x10, 30x25	English	News	Both	
DUC 2004 [106]	100x10	English, Arabic	News	Both	
DUC 2005 [107]	50x32	English	News	Multi	
DUC 2006 [108]	50x25	English	News	Multi	
DUC 2007 [108]	25x10	English	News	Multi	
Gigaword 5 [109]	9876086	English	News	Single	https://catalog.ldc.upenn.edu/LDC2011T07
CNN/Daily Mail [110]	312084	English	News	Single	https://github.com/deepmind/rc-data/
Pubmed [111]	19717	English	Scientific Publications	Multi	https://paperswithcode.com/dataset/pubmed
INDOSUM [112]	Upto 200x10	Indonesian	News	Both	https://github.com/kata-ai/indosum
EASC [113]	153	Arabic	News, Wekipedia	Single	https://www.lancaster.ac.uk/staff/elhaj/corpora.html
SummBank [114]	40x10	English, Chines	News	Both	https://catalog.ldc.upenn.edu/LDC2003T16
Opinosis [115]	51x100	English	Reviews	Multi	http://kavita-ganesan.com/opinosis-opinion-dataset/
LCSTS [116]	2400591	Chines	Blogs	Single	http://icrc.hitsz.edu.cn/Article/show/139.html
CAST [117]	147	English	News	Single	http://clg.wlv.ac.uk/projects/CAST/corpus/index.php

XII. AUTOMATIC TEXT SUMMARIZATION EVALUATIONS MEASURES

These are some of the most regularly used NLP metrics, and they may also be used to assess text summarization systems to measure the effectiveness of ATS systems. Among them, some of the standard measures are:

A. RECALL-ORIENTED UNDERSTUDY FOR GISTING EVALUATION (ROUGE)

This group of metrics assess the degree to which the reference summary and the system-generated summary overlap [38], [48], [49]. ROUGE-L measures the longest common subsequence, whereas ROUGE-N measures the n-gram overlap between the two summaries.

$$ROUGE - N = \frac{\sum s \in Summary \sum n \in N - grams\ CountMatch(n, s)}{\sum s \in Summary \sum n{\in}N - grams\ Count(n)} \tag{1}$$

$$ROUGE - L = \frac{\sum s \in Summary\ LCS(s, Reference)}{\sum s \in Summary\ |s|} \tag{2}$$

where, CountMatch(n,s) counts the number of times an N-gram n appears in the summary s. N-grams are contiguous sequences of N words. Count(n) counts the total number of N-grams in the reference summary. LCS(s, Reference) is the length of the longest common subsequence between the summary s and the reference summary.

B. BILINGUAL EVALUATION UNDERSTUDY (BLEU)

This measure may be used for text summarization as well as machine translation, however, it is most frequently employed for translation [8], [48], [49]. Calculating the n-gram overlap determines how similar the system-generated summary and the reference summary are:

$$BLUE = BP.\exp\left(\sum_{n=1}^{N}\frac{1}{N}logPn\right) \tag{3}$$

where: Pn is the modified n-gram precision (the ratio of the number of N-grams that appear in the system summary and the reference summary). N-grams are contiguous sequences of N words. BP is the brevity penalty factor that penalizes summaries that are too short.

C. METRIC FOR EVALUATION OF TRANSLATION WITH EXPLICIT ORDERING (METER)

The similarity between the summary produced by the system and the reference summary is determined using this metric, which weighs accuracy, recall, and alignment [8].

$$METEOR = P_{align}.R_{align}.F_{mean}.IDF \tag{4}$$

where: Ralign is the recall of the alignment between the system summary and the reference summary. Palign is the precision of the alignment between the system summary and the reference summary. Fmean is the mean of precision and recall. IDF is the inverse document frequency that weights rare words more heavily.

D. PRECISION

Precision is defined as the fraction of relevant picked elements [73], [118]. Precision may be calculated in the context of text summarization as follows (5), shown at the bottom of the page.

E. RECALL

The proportion of relevant items chosen is measured by recall [119], [120]. Recall may be calculated in the context of text summary as follows (6), shown at the bottom of the page.

F. F1-SCORE

The F1-Score is the harmonic mean of accuracy and recall, and it gives a balanced evaluation of the trade-off between the two [119], [121]. The F1-Score equation is as follows:

$$F1 - Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{7}$$

G. ACCURACY

Accuracy is defined as the fraction of correctly categorized objects (in this example, summaries) [122], [123]. The accuracy equation is as follows:

$$Accuracy = \frac{Number\ of\ correct\ system\ summaries}{Total\ number\ of\ system\ summaries} \tag{8}$$

XIII. CONCLUSION

ATS reduces the size of a source text while preserving its informational value and overall meaning. Due to the abundance of information we get and the development of Internet technology, ATS has developed into an effective method for assessing text data. A well-known problem in NLP is the automated summary of the text. ATS is a fascinating study field with a wealth of potential applications. This comprehensive survey paper has provided a thorough overview of ATS, covering various challenges, types, classifications, approaches, applications, methods, implementations, processing and preprocessing, linguistic analysis, datasets, and evaluation measures. Despite the enormous advancements in ATS research, there are still several issues that need to be resolved, such as increasing the summaries' accuracy and dealing with terminology peculiar to certain domains. Several problems consistently have an influence on the degree of summary accuracy in ATS. Content selection continues to be a major challenge since ATS systems must precisely identify and prioritize essential information while removing redundant data. A thorough comprehension of context and semantics is necessary to ensure accuracy while producing summaries that are human-like using abstractive summarization. Another ongoing problem is how to deal with lexical and contextual ambiguities in language. The pursuit for improved summary accuracy in ATS is impacted by a number of ongoing issues, including adapting ATS to different domains with domain-specific terminology and writing styles, maintaining coherence and fluency in summaries, addressing unpredictability in source texts, and generating strong assessment measures. The way we engage with textual data has a lot of opportunities to change, in our opinion.

$$Precision = \frac{Number\ of\ overlapping\ words\ in\ system\ and\ reference\ summary}{Total\ number\ of\ words\ in\ system\ summary} \tag{5}$$

$$Recall = \frac{Number\ of\ overlapping\ words\ in\ system\ and\ reference\ summary}{Total\ number\ of\ words\ in\ reference\ summary} \tag{6}$$

The future of this discipline is defined by both enormous promise and problems as we get to the end of our review of ATS approaches. Large-scale pre-trained language models like GPT-3 enable the processing of multimodal data while also providing customization options for specialized topics and languages. The highest priority will be given to ethical issues, including prejudice mitigation and responsible usage. Key areas of study include real-time summarization, enhanced assessment measures, and multidisciplinary cooperation. Text summarization is at the confluence of AI innovation and ethical responsibility in this rapidly changing environment, necessitating a careful balancing act between maximizing the potential of cutting-edge tools and guaranteeing fairness, inclusivity, and utility across diverse user needs and industry demands. As AI technologies proliferate, urgent ethical questions arise that demand effective steps for bias prevention, transparency, and responsible usage. Developing real-time summary systems, improving assessment criteria to match sophisticated abstractive summarization, and encouraging multidisciplinary cooperation among NLP professionals, subject-matter experts, and ethicists are some of the unmet difficulties.

Looking forward, we suggest several future research directions in ATS, such as developing more efficient models that can take into account external knowledge and contextual data, developing reliable methods for handling large volumes of data and investigating new approaches for summarizing various types of media, such as images and videos. Overall, ATS has a great deal of potential to transform how we work with textual data. We hope that our survey will help researchers interested in this discipline and inspire more advancement in this fascinating field of study.

REFERENCES

[1] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," *Expert Syst. Appl.*, vol. 172, Jun. 2021, Art. no. 114652, doi: 10.1016/j.eswa.2021.114652.

[2] G. K. Kumar and D. M. Rani, "Paragraph summarization based on word frequency using NLP techniques," *AIP Conf. Proc.*, vol. 2317, Feb. 2021, Art. no. 060001, doi: 10.1063/5.0037283.

[3] A. C. Sanders, R. C. White, L. S. Severson, R. Ma, R. McQueen, H. C. A. Paulo, Y. Zhang, J. S. Erickson, and K. P. Bennett, "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse," *AMIA J. Summits Transl. Sci. Proc.*, vol. 2021, pp. 555–564, May 2021.

[4] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Develop.*, vol. 2, no. 2, pp. 159–165, Apr. 1958.

[5] E. Hovy and C.-Y. Lin, "Automated text summarization and the SUMMARIST system," in *Proc. Tipster Text Program Phase III*, 1996, pp. 197–214, doi: 10.3115/1119089.1119121.

[6] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: A brief survey," 2017, *arXiv:1707.02268*.

[7] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: A survey," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 1–66, Jan. 2017.

[8] A. A. Syed, F. L. Gaol, and T. Matsuo, "A survey of the state-of-the-art models in neural abstractive text summarization," *IEEE Access*, vol. 9, pp. 13248–13265, 2021, doi: 10.1109/ACCESS.2021.3052783.

[9] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113679, doi: 10.1016/j.eswa.2020.113679.

[10] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitzt, "Multi-document summarization by sentence extraction," in *Proc. ANLP/NAACL Workshop Autom. Summarization*, 1998, pp. 40–48.

[11] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," *Expert Syst. Appl.*, vol. 120, pp. 43–56, Apr. 2019, doi: 10.1016/j.eswa.2018.11.022.

[12] M. Tomer and M. Kumar, "Multi-document extractive text summarization based on firefly algorithm," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6057–6065, Sep. 2022, doi: 10.1016/j.jksuci.2021.04.004.

[13] R. Rautray and R. C. Balabantaray, "An evolutionary framework for multi document summarization using cuckoo search approach: MDSCSA," *Appl. Comput. Informat.*, vol. 14, no. 2, pp. 134–144, Jul. 2018, doi: 10.1016/j.aci.2017.05.003.

[14] K. Jezek and J. Steinberger, "Automatic text summarization: (The state of the art 2007 and new challenges)," in *Proc. Znalosti*, 2008, pp. 1–12.

[15] P. Mishra. *Problems with Existing Abstractive Text Summarization Models—Even SOTA*. Accessed: Apr. 4, 2022. [Online]. Available: https://towardsdatascience.com/entity-level-factual-consistency-in-abstractive-text-summarization-cb19e8a48397

[16] A. Shelton, C. J. Lemons, and J. Wexler, "Supporting main idea identification and text summarization in middle school co-taught classes," *Intervent School Clinic*, vol. 56, no. 4, pp. 217–223, Mar. 2021, doi: 10.1177/1053451220944380.

[17] J. Steinberger and K. Je, "Text summarization: An old challenge and new approaches," in *Foundations of Computational, Intelligence*, vol. 6. Berlin, Germany: Springer, 2009, pp. 127–149.

[18] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.

[19] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in *Proc. ACL Interact. Poster Demonstration Sessions*, 2004, pp. 170–173.

[20] D. R. Radev, S. Blair-Goldensohn, and Z. Zhang, "Experiments in single and multi-document summarization using MEAD," in *Proc. 1st Document Understand. Conf.*, 2001, pp. 1–8.

[21] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Jan. 2017, pp. 1–6.

[22] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258–268, Aug. 2010.

[23] C. Ma, W. Emma Zhang, M. Guo, H. Wang, and Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," 2020, *arXiv:2011.04843*.

[24] M. J. Mohan, C. Sunitha, A. Ganesh, and A. Jaya, "A study on ontology based abstractive summarization," *Proc. Comput. Sci.*, vol. 87, pp. 32–37, 2016, doi: 10.1016/j.procs.2016.05.122.

[25] I. K. Bhat, M. Mohd, and R. Hashmy, "SumItUp: A hybrid single-document text summarizer," in *Soft Computing: Theories and Applications* (Advances in Intelligent Systems and Computing), vol. 583. Singapore: Springer, Apr. 2018, pp. 619–634, doi: 10.1007/978-981-10-5687-1_56.

[26] S. Mohammed, "Introducing the new JETWI associate editor-in-chief," *J. Emerg. Technol. Web Intell.*, vol. 5, no. 1, p. 1, Feb. 2013, doi: 10.4304/jetwi.5.1.1.

[27] M. Joshi, H. Wang, and S. McClean, "Dense semantic graph and its application in single document summarisation," in *Emerging Ideas on Information Filtering and Retrieval*. Cham, Switzerland: Springer, 2018, pp. 55–67.

[28] V. K. Gupta and T. J. Siddiqui, "Multi-document summarization using sentence clustering," in *Proc. 4th Int. Conf. Intell. Human Comput. Interact. (IHCI)*, Dec. 2012, pp. 1–5.

[29] J. K. Yogan, O. S. Goh, B. Halizah, H. C. Ngo, and C. Puspalata, "A review on automatic text summarization approaches," *J. Comput. Sci.*, vol. 12, no. 4, pp. 178–190, 2016.

[30] D. Sahoo, R. Balabantaray, M. Phukon, and S. Saikia, "Aspect based multi-document summarization," in *Proc. Int. Conf. Comput., Commun. Autom. (ICCCA)*, Apr. 2016, pp. 873–877.

[31] M. Mohd, R. Jan, and M. Shah, "Text document summarization using word embedding," *Expert Syst. Appl.*, vol. 143, Apr. 2020, Art. no. 112958.

[32] K. Takeuchi, "A study on operations used in text summarization," Ph.D. thesis, Nara Inst. Sci. Technol., 2002, Paper DT9961016.

[33] F. Dernoncourt, M. Ghassemi, and W. Chang, "A repository of corpora for summarization," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2019, pp. 3221–3227.

[34] K. Woodsend and M. Lapata, "Automatic generation of story highlights," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2010, pp. 565–574.

[35] A. Khan, N. Salim, H. Farman, M. Khan, B. Jan, A. Ahmad, I. Ahmed, and A. Paul, "Abstractive text summarization based on improved semantic graph approach," *Int. J. Parallel Program.*, vol. 46, no. 5, pp. 992–1016, Oct. 2018.

[36] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive text summarization: Enhancing sequence-to-sequence models using word sense disambiguation and semantic content generalization," *Comput. Linguistics*, vol. 47, no. 4, pp. 813–859, Dec. 2021.

[37] D. Supreetha, S. B. Rajeshwari, and J. S. Kallimani, "Abstractive text summarization," *J. Xidian Univ.*, vol. 14, no. 6, pp. 26884–26888, 2020, doi: 10.37896/jxu14.6/094.

[38] G. Sharma and D. Sharma, "Automatic text summarization methods: A comprehensive review," *Social Netw. Comput. Sci.*, vol. 4, no. 1, p. 33, 2023, doi: 10.1007/s42979-022-01446-w.

[39] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Syst. Appl.*, vol. 121, pp. 49–65, May 2019, doi: 10.1016/J.ESWA.2018.12.011.

[40] N. Moratanch and S. Chitrakala, "Anaphora resolved abstractive text summarization (AR-ATS) system," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 4569–4597, Jan. 2023.

[41] A. Khan and N. Salim, "A review on abstractive summarization methods," *J. Theor. Appl. Inf. Technol.*, vol. 59, no. 1, pp. 64–72, 2014.

[42] J. N. Madhuri and R. Ganesh Kumar, "Extractive text summarization using sentence ranking," in *Proc. Int. Conf. Data Sci. Commun. (IconDSC)*, Mar. 2019, pp. 1–3.

[43] A. Jain, D. Bhatia, and M. K. Thakur, "Extractive text summarization using word vector embedding," in *Proc. Int. Conf. Mach. Learn. Data Sci. (MLDS)*, Dec. 2017, pp. 51–55.

[44] S. Alshaina, A. John, and A. G. Nath, "Multi-document abstractive summarization based on predicate argument structure," in *Proc. IEEE Int. Conf. Signal Process., Informat., Commun. Energy Syst. (SPICES)*, Aug. 2017, pp. 1–6.

[45] R. Nallapati, B. Zhou, C. N. dos santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," 2016, *arXiv:1602.06023*.

[46] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," 2017, *arXiv:1704.04368*.

[47] S. Ma, X. Sun, J. Lin, and X. Ren, "A hierarchical end-to-end model for jointly improving text summarization and sentiment classification," 2018, *arXiv:1805.01089*.

[48] A. K. Yadav, A. Singh, M. Dhiman, Vineet, R. Kaundal, A. Verma, and D. Yadav, "Extractive text summarization using deep learning approach," *Int. J. Inf. Technol.*, vol. 14, no. 5, pp. 2407–2415, Aug. 2022, doi: 10.1007/s41870-022-00863-7.

[49] A. Elsaid, A. Mohammed, L. F. Ibrahim, and M. M. Sakre, "A comprehensive review of Arabic text summarization," *IEEE Access*, vol. 10, pp. 38012–38030, 2022, doi: 10.1109/ACCESS.2022.3163292.

[50] P. Agarwal and S. Mehta, "Empirical analysis of five nature-inspired algorithms on real parameter optimization problems," *Artif. Intell. Rev.*, vol. 50, no. 3, pp. 383–439, Oct. 2018.

[51] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 43–76.

[52] P. Mehta and P. Majumder, "Effective aggregation of various summarization techniques," *Inf. Process. Manage.*, vol. 54, no. 2, pp. 145–158, Mar. 2018.

[53] N. Nazari and M. Mahdavi, "A survey on automatic text summarization," *J. AI Data Mining*, vol. 7, no. 1, pp. 121–135, 2019.

[54] S. Wang, X. Zhao, B. Li, B. Ge, and D. Tang, "Integrating extractive and abstractive models for long text summarization," in *Proc. IEEE Int. Congr. Big Data (BigData Congress)*, Jun. 2017, pp. 305–312.

[55] K. Al-Sabahi, Z. Zhang, J. Long, and K. Alwesabi, "An enhanced latent semantic analysis approach for Arabic document summarization," *Arabian J. Sci. Eng.*, vol. 43, no. 12, pp. 8079–8094, Dec. 2018.

[56] M. Mohamed and M. Oussalah, "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1356–1372, Jul. 2019.

[57] H. Kobayashi, M. Noguchi, and T. Yatsuka, "Summarization based on embedding distributions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1984–1989.

[58] L. Chen and M. L. Nguyen, "Sentence selective neural extractive summarization with reinforcement learning," in *Proc. 11th Int. Conf. Knowl. Syst. Eng. (KSE)*, Oct. 2019, pp. 1–5.

[59] J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, "Experimental analysis of multiple criteria for extractive multi-document text summarization," *Expert Syst. Appl.*, vol. 140, Feb. 2020, Art. no. 112904.

[60] A. Sharaff, A. S. Khaire, and D. Sharma, "Analysing fuzzy based approach for extractive text summarization," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 906–910.

[61] I. K. Bhat, M. Mohd, and R. Hashmy, "SumItUp: A hybrid single-document text summarizer," in *Proc. Soft Comput., Theories Appl. (SoCTA)*, vol. 1. Singapore: Springer, 2018, pp. 619–634.

[62] E. Lloret, M. T. Romá-Ferri, and M. Palomar, "COMPENDIUM: A text summarization system for generating abstracts of research papers," *Data Knowl. Eng.*, vol. 88, pp. 164–175, Nov. 2013.

[63] A. P. Patil, S. Dalmia, S. A. A. Ansari, T. Aul, and V. Bhatnagar, "Automatic text summarizer," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 1530–1534.

[64] H. Wang, K. Qin, R. Y. Zakari, G. Lu, and J. Yin, "Deep neural network-based relation extraction: An overview," *Neural Comput. Appl.*, vol. 34, no. 6, pp. 4781–4801, Mar. 2022.

[65] H. Gupta and M. Patel, "Method of text summarization using LSA and sentence based topic modelling with bert," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 511–517.

[66] A. K. Das, B. Thumu, A. Sarkar, S. Vimal, and A. K. Das, "Graph-based text summarization and its application on COVID-19 Twitter data," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 30, no. 3, pp. 513–540, Jun. 2022.

[67] B. D. Lund and T. Wang, "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?" *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, Jan. 2023.

[68] Z. Huang and Z. Xie, "A patent keywords extraction method using TextRank model with prior public knowledge," *Complex Intell. Syst.*, vol. 8, no. 1, pp. 1–12, Feb. 2022.

[69] H. Gupta and M. Patel, "Study of extractive text summarizer using the Elmo embedding," in *Proc. 4th Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Oct. 2020, pp. 829–834.

[70] M. Wang, J. He, and P. R. Hahn, "Local Gaussian process extrapolation for BART models with applications to causal inference," 2022, *arXiv:2204.10963*.

[71] O. Zheng, M. Abdel-Aty, D. Wang, Z. Wang, and S. Ding, "ChatGPT is on the Horizon: Could a large language model be all we need for intelligent transportation?" 2023, *arXiv:2303.05382*.

[72] N. Jiwani, K. Gupta, and P. Whig, "Analysis of the potential impact of omicron crises using NLTK (natural language toolkit)," in *Proc. 3rd Doctoral Symp. Comput. Intell. (DoSCI)*, 2022, pp. 445–454.

[73] H. N. Serere, B. Resch, and C. R. Havas, "Enhanced geocoding precision for location inference of tweet text using spaCy, nominatim and Google Maps. A comparative analysis of the influence of data selection," *PLoS ONE*, vol. 18, no. 3, Mar. 2023, Art. no. e0282942.

[74] F. K. Sufi and I. Khalil, "Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis," *IEEE Trans. Computat. Social Syst.*, early access, Mar. 18, 2022, doi: 10.1109/TCSS.2022.3157142.

[75] W. Jiang, N. Synovic, M. Hyatt, T. R. Schorlemmer, R. Sethi, Y.-H. Lu, G. K. Thiruvathukal, and J. C. Davis, "An empirical study of pre-trained model reuse in the hugging face deep learning model registry," 2023, *arXiv:2303.02552*.

[76] A. Esuli and F. Sebastiani, "Training data cleaning for text classification," in *Proc. Conf. Theory Inf. Retr.*, 2009, pp. 29–41.

[77] I. Pak and P. L. Teh, "Text segmentation techniques: A critical review," *Innovative Computing, Optimization and Its Applications: Modeling and Simulations*. Cham, Switzerland: Springer, 2018, pp. 167–181.

[78] J. Kaur and P. K. Buttar, ''A systematic review on stopword removal algorithms,'' *Int. J. Future Revolution Comput. Sci. Commun. Eng.*, vol. 4, no. 4, pp. 207–210, 2018.

[79] R. Pramana, Debora, J. J. Subroto, A. A. S. Gunawan, and Anderies, ''Systematic literature review of stemming and lemmatization performance for sentence similarity,'' in *Proc. IEEE 7th Int. Conf. Inf. Technol. Digit. Appl. (ICITDA)*, Nov. 2022, pp. 1–6.

[80] I. Budi and R. R. Suryono, ''Application of named entity recognition method for Indonesian datasets: A review,'' *Bull. Electr. Eng. Informat.*, vol. 12, no. 2, pp. 969–978, Apr. 2023.

[81] S. G. Kanakaraddi and S. S. Nandyal, ''Survey on parts of speech tagger techniques,'' in *Proc. Int. Conf. Current Trends Towards Converging Technol. (ICCTCT)*, Mar. 2018, pp. 1–6.

[82] H. Kaur, V. Mangat, and Nidhi, ''A survey of sentiment analysis techniques,'' in *Proc. Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Feb. 2017, pp. 921–925.

[83] R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, ''Anaphora and coreference resolution: A review,'' *Inf. Fusion*, vol. 59, pp. 139–162, Jul. 2020.

[84] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni, ''Natural Language Processing (NLP) based text summarization—A survey,'' in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1310–1317.

[85] T. Boroş, Ş. D. Dumitrescu, and R. Burtica, ''NLP-Cube: End-to-end raw text processing with neural networks,'' in *Proc. CoNLL Shared Task, Multilingual Parsing Raw Text Universal Dependencies*, 2018, pp. 171–179.

[86] A. F. T. Martins and N. A. Smith, ''Summarization with a joint model for sentence extraction and compression,'' in *Proc. Workshop Integer Linear Program. Natural Langauge Process. (ILP)*, 2009, pp. 1–9.

[87] T. Gupta, ''Keyword extraction: A review,'' *Int. J. Eng. Appl. Sci. Technol.*, vol. 2, no. 4, pp. 215–220, 2017.

[88] Y. Li, T. Fei, and F. Zhang, ''A regionalization method for clustering and partitioning based on trajectories from NLP perspective,'' *Int. J. Geographical Inf. Sci.*, vol. 33, no. 12, pp. 2385–2405, Dec. 2019.

[89] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Sun, B. Xu, and Z. Zhao, ''Neural network-based approaches for biomedical relation classification: A review,'' *J. Biomed. Informat.*, vol. 99, Nov. 2019, Art. no. 103294.

[90] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, ''Multi-document summarization via deep learning techniques: A survey,'' *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–37, May 2023.

[91] P. G. Magdum and S. Rathi, ''A survey on deep learning-based automatic text summarization models,'' in *Proc. Adv. Artif. Intell. Data Eng. (AIDE)*, 2021, pp. 377–392.

[92] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, ''A survey on contrastive self-supervised learning,'' *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[93] N. Jung and G. Lee, ''Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning,'' *Adv. Eng. Informat.*, vol. 41, Aug. 2019, Art. no. 100917.

[94] W. Y. Wang, J. Li, and X. He, ''Deep reinforcement learning for NLP,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, Tutorial Abstr.*, 2018, pp. 19–21.

[95] E. Strubell, A. Ganesh, and A. McCallum, ''Energy and policy considerations for deep learning in NLP,'' 2019, *arXiv:1906.02243*.

[96] A. Chiche and B. Yitagesu, ''Part of speech tagging: A systematic review of deep learning and machine learning approaches,'' *J. Big Data*, vol. 9, no. 1, pp. 1–25, Jan. 2022.

[97] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, ''Named entity recognition approaches and their comparison for custom NER model,'' *Sci. Technol. Libraries*, vol. 39, no. 3, pp. 324–337, Jul. 2020.

[98] A. Doitch, R. Yazdi, T. Hazan, and R. Reichart, ''Perturbation based learning for structured NLP tasks with application to dependency parsing,'' *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 643–659, Nov. 2019.

[99] M. R. Hasan, M. Maliha, and M. Arifuzzaman, ''Sentiment analysis with NLP on Twitter data,'' in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME2)*, Jul. 2019, pp. 1–4.

[100] N. Clarke, P. Foltz, and P. Garrard, ''How to do things with (thousands of) words: Computational approaches to discourse analysis in Alzheimer's disease,'' *Cortex*, vol. 129, pp. 446–463, Aug. 2020.

[101] D. Gillick, B. Favre, and D. Hakkani-Tür, ''The ICSI summarization system at TAC 2008,'' Int. Comput. Sci. Inst., Berkeley, USA, Tech. Rep., 2008.

[102] C. Long, M. Huang, and X. Zhu, ''Tsinghua University at TAC 2009: Summarizing multi-documents by information distance,'' Tsinghua Univ., China, Tech. Rep., 2009.

[103] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, ''Overview of the TAC 2010 knowledge base population track,'' in *Proc. 3rd Text Anal. Conf. (TAC)*, 2010, vol. 3, no. 2, p. 3.

[104] H. Oufaida, O. Nouali, and P. Blache, ''Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization,'' *J. King Saud Univ. Comput. Inf. Sci.*, vol. 26, no. 4, pp. 450–461, Dec. 2014.

[105] Y. Gallina, F. Boudin, and B. Daille, ''KPTimes: A large-scale dataset for keyphrase generation on news documents,'' 2019, *arXiv:1911.12559*.

[106] Y. Zhang, J. E. Meng, and M. Pratama, ''Extractive document summarization based on convolutional neural networks,'' in *Proc. 42nd Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2016, pp. 918–922.

[107] R. Passonneau, ''Evaluating an evaluation method: The pyramid method applied to 2003 document understanding conference (DUC) data,'' Columbia Univ., New York, NY, USA, Tech. Rep., 2006.

[108] Y. Seki, K. Eguchi, N. Kando, and M. Aono, ''Opinion-focused summarization and its analysis at DUC 2006,'' in *Proc. Document Understand. Conf. (DUC)*, 2006, pp. 122–130.

[109] S. R. Eide, N. Tahmasebi, and L. Borin, ''The Swedish culturomics gigaword corpus: A one billion word Swedish reference dataset for NLP,'' in *Proc. From Digitization Knowl. Workshop*, 2016, pp. 8–12.

[110] M. Grusky, M. Naaman, and Y. Artzi, ''Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies,'' 2018, *arXiv:1804.11283*.

[111] V. Gupta, P. Bharti, P. Nokhiz, and H. Karnick, ''SumPubMed: Summarization dataset of PubMed scientific articles,'' in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., Student Res. Workshop*, 2021, pp. 292–303.

[112] N. Khotimah and A. S. Girsang, ''Indonesian news articles summarization using genetic algorithm,'' *Eng. Lett.*, vol. 30, no. 1, pp. 1–9, 2022.

[113] D. Suleiman and A. A. Awajan, ''Deep learning based extractive text summarization: Approaches, datasets and evaluation measures,'' in *Proc. 6th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Oct. 2019, pp. 204–210.

[114] H. Zhang and J. Wang, ''An unsupervised semantic sentence ranking scheme for text documents,'' *Integr. Comput.-Aided Eng.*, vol. 28, no. 1, pp. 17–33, Dec. 2020.

[115] R. C. Belwal, S. Rai, and A. Gupta, ''Text summarization using topic-based vector space model and semantic measure,'' *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102536.

[116] B. Hu, Q. Chen, and F. Zhu, ''LCSTS: A large scale Chinese short text summarization dataset,'' 2015, *arXiv:1506.05865*.

[117] J. Dalton, C. Xiong, V. Kumar, and J. Callan, ''CAsT-19: A dataset for conversational information seeking,'' in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2020, pp. 1985–1988.

[118] R. Naseem, B. Khan, M. A. Shah, K. Wakil, A. Khan, W. Alosaimi, M. I. Uddin, and B. Alouffi, ''Performance assessment of classification algorithms on early detection of liver syndrome,'' *J. Healthcare Eng.*, vol. 2020, Dec. 2020, Art. no. 6680002, doi: 10.1155/2020/6680002.

[119] R. Naseem, B. Khan, A. Ahmad, A. Almogren, S. Jabeen, B. Hayat, and M. A. Shah, ''Investigating tree family machine learning techniques for a predictive system to unveil software defects,'' *Complexity*, vol. 2020, Nov. 2020, Art. no. 6688075, doi: 10.1155/2020/6688075.

[120] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, ''An empirical evaluation of machine learning techniques for chronic kidney disease prophecy,'' *IEEE Access*, vol. 8, pp. 55012–55022, 2020, doi: 10.1109/ACCESS.2020.2981689.

[121] S. Ouni, F. Fkih, and M. N. Omri, ''Toward a new approach to author profiling based on the extraction of statistical features,'' *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–16, Dec. 2021, doi: 10.1007/s13278-021-00768-6.

[122] Y. Hayashi and K. Fukunaga, ''Accuracy of rule extraction using a recursive-rule extraction algorithm with continuous attributes combined with a sampling selection technique for the diagnosis of liver disease,'' *Informat. Med. Unlocked*, vol. 5, pp. 26–38, 2016, doi: 10.1016/j.imu.2016.10.001.

[123] S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P. K. Singh, and W.-C. Hong, "Machine learning adoption in blockchain-based smart applications: The challenges, and a way forward," *IEEE Access*, vol. 8, pp. 474–488, 2020, doi: 10.1109/ACCESS.2019.2961372.

BILAL KHAN received the M.C.S. degree from Northern University, Nowshera, and the M.S.C.S. degree in computer science from the City University of Science and Information Technology (CUSIT), Peshawar, Pakistan. He is currently pursuing the Ph.D. degree with the University of Engineering and Technology Mardan. He is a Lecturer with the Department of Computer Science, CUSIT. His research interests include natural language processing, machine learning, data science, software engineering, and bio-informatics.

ZOHAIB ALI SHAH received the B.S. degree in computer systems from the University of Engineering and Technology (UET) Mardan, Pakistan, in 2022, where he is currently pursuing the master's degree in computer software engineering. He is a Laboratory Engineer with the Department of Computer Software Engineering, MCS, NUST, and a Teacher Assistant with UET Mardan. His research interests include artificial intelligence and machine learning. He received the Best Student of the Year Award in Matric and the 12th position in the General Science Group in Intermediate at Mardan Board. He also received the Dean's Honor Award for outstanding academic results (2020–2021 and 2021–2022).

MUHAMMAD USMAN (Senior Member, IEEE) received the B.Sc. degree in computer information systems engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2004, the M.Sc. degree in computer engineering from the Center of Advanced Studies in Engineering (CASE), Islamabad, Pakistan, in 2007, and the Ph.D. degree from the University of Ulsan, South Korea, in 2016. He is currently a Professor with the Department of Computer Science, UET Mardan. His research interests include security and energy efficiency in wireless networks, software engineering, and machine learning. In 2015, he received the Best SCI(E) Paper Award from the Korean Government through the BK21+ project. He is a Reviewer of various reputable international journals, such as IEEE Transactions on Wireless Communications, IEEE Access, IEEE Transactions on Cognitive Communication and Networking, IEEE Sensors Journal, IEEE Systems Journal, *Applied Mathematics & Information Sciences*, and *The Computer Journal*.

INAYAT KHAN received the Ph.D. degree in computer science from the Department of Computer Science, University of Peshawar, Pakistan. He is currently an Assistant Professor of computer science with the University of Engineering and Technology Mardan, Pakistan. He has published more than 60 research articles. His research interests include lifelogging, healthcare, deep learning, ubiquitous computing, accessibility, and mobile-based assistive systems for people with disabilities. He is a reviewer of various reputable international journals.

BADAM NIAZI received the M.S. degree from the Department of Computer Science, University of Peshawar. He is currently an Assistant Professor of computer science with the Faculty of Computing, Nangarhar University, Jalalabad, Pakistan. He has published several research papers in well reputed international journals and conferences. His research interest includes mobile-based assistive technologies for the special with special needs.