A Comprehensive Survey on Text Summarization for Indian Languages: Opportunities, Challenges and Future Prospects

[1]Shilpa Serasiya, [2]Uttam Chauhan

[1] *Research Scholar, Computer Engineering, Gujarat Technological University, Ahmedabad, India*
shilpapatel84@gmail.com
[2] *Assistant Professor, Computer Engineering Department, Vishwakarma Government Engineering College, Ahmedabad, India*
ugchauhan@vgecg.ac.in

KEYWORDS
Extractive, Abstractive, Text summarization, Deep Learning, Indian Regional Languages

ABSTRACT
An increasing quantum of data available on the web, news websites, published articles in various fields of study, and electronic books have generated a valuable resource for extracting and analyzing information. The main challenge for researchers has been that of accessing accurate and reliable data. This information must be summarized to retrieve helpful knowledge within a reasonable period. Text summarization is a crucial Natural Language Processing (NLP) task that aims to condense lengthy documents into shorter, coherent summaries while retaining the essential information. Text summarization is divided into Extractive and Abstractive Summarization. The extractive summarizer extracts the basic sentences or phrases from the original document. In contrast, an Abstractive summarizer generate a summary by rephrasing the original text with new one which is closed to the human-made. With the increasing availability of textual data in multiple Indian languages, effective summarization techniques are required to facilitate content understanding, especially for non-English users. Indian languages, such as Hindi, Tamil, Bengali, Gujarati, and Marathi etc. have complex linguistic structures, making summarization challenging. Countable research works is carried out with extractive methods for Indian language, now researcher move towards the abstractive summarization. This paper presents an overview of summarization techniques, datasets, and evaluation metrics for Indian languages. In addition, the survey represents the Deep Learning-based text summarization with valuable adoption of conventional approaches to uplift the abstractive text summarization.

1. INTRODUCTION

In this century, our world is parachuted by gathering and broadcasting a massive amount of data. Propelled by modern technological innovations, As per International Data Corporation (IDC), the amount of digital data circulating annually around the world emerges from 4.4 Zettabytes in 2013-14 to 190 Zettabytes in 2025-26. With an increasing quantum of data available on the web, increasing news websites, numbers of published articles in various fields of study, and publications of different electronic books. It's tough for a normal human being to summarize manually large documents of text. Thus, the main challenge for the researcher in this century has been retrieving important information from the text quickly within a short period. Even the solution to the problem in evaluating the summary is a big problem to resolve.[35].

Considering this problem, text summarization has become essential to preserve. It is the process of automatically creating and condensing the form of a given document and maintaining its source information into a shorter version with overall meaning. The summarization technique is convenient in various fields such as summary of online news articles, previews are produced as snippets in search engines and product review summaries, automated research abstracts, one-line email summaries, for government officials, business organizations, abstracted information summaries, etc. with the minimum human involvement [4].

Natural language is one of the challenging problems in language Processing, while summarization for any regional language is a crucial part of that. Humans can select salient words and summarize any segment or text. But in it is difficult for the computer system. Because summary generation is involved with language awareness, interpretation, and presentation of common-sense knowledge like humans.[23] As we know that text summarization is the process by which the document is reduced to few sentences without changing the

gist of the document. Text summarization is broadly categorized into different types based on how summaries are generated and the method used. Here are some common categorizations of text summarization based on the Input type, purpose, and output type etc. Table 1 describe the categories of text summarization.

Table 1: Category of text summarization

	Type	Description	Example Methods
Based on Content Type	Extractive	Selects key sentences from the text	TextRank, LexRank, BERT
	Abstractive	Generates new sentences preserving meaning	Seq2Seq, Transformer, GPT, BART
Based on Input Type	Single-Document	Summarizes one document	Used in news summarization
	Multi-Document	Summarizes multiple related documents	Used in research reviews, topic-based summaries
Based on Output Type	Indicative	Gives an overview	Research paper abstracts
	Informative	Provides detailed key information	Legal, medical, financial reports
Based on Purpose Type	Generic	Consider entire information useful	TF-IDF, TextRank, BERTSUM, LSA
	Query-based	Required specific information from the original document	TF-IDF, BM25
Based on Techniques	Statistical/Graph-Based	Uses TF-IDF, TextRank	TextRank, LexRank, TF-IDF
	Machine Learning-Based	Uses ML models for ranking sentences	Random Forest, SVM, BERT
	Deep Learning-Based	Uses neural networks for summary generation	Transformer, GPT, BART, T5

Out of the above five types of text summarization, extractive and abstractive summarization is very popular and so most researchers work for this summarization.

Extractive summarization selects key sentences directly from the original text based on their importance. It uses statistical methods such as title/headline, TF-IDF, sentence length, cue method, proper noun, similarity etc., Graph-based model such as TextRank, PageRank and machine learning approaches such as support vector machines (SVM), decision trees, random forest to find the sentences required for generating a summary without changing the original document. Most research works are carried out in extractive methods [16,27,36]. **Abstractive** text summarization(ATS) understands the document and rephrases the input text with new one to generate a summary which is closed to the human-made. So, it requires natural language processing (NLP) to understand the document and advanced machine learning techniques for summarization [18,37].

The summary must be compact while conveying important information about the original text; before summarizing the text, some pre-processing tasks should be performed: Sentence boundary identification (sentence position, word position), stop word removal, and stemming. [33] Structured and semantic based approaches are used in abstractive summarization. Cognitive schemas like templates, rules extraction, tree, ontology, lead and body phrase are used to take out the most important features/information. At the same time, semantic-based approaches are more anxious with the semantics of the text. Semantic-based strategies include the multi-modal semantic method, information item method, and semantic graph-based method[38]. Challenges of extractive text summarization includes ccoherence & readability, lack of paraphrasing, dependency on sentence selection while abstractive summarization includes language generation complexity, risk of hallucination and computational cost.

Deep Learning-based abstractive text summarization with valuable adoption of conventional approaches to uplift the abstractive text summarization that mathods includes Seq2Seq Models using LSTM, GRU-based architectures, mT5, mBART, BERTSUM, GPT-3 of transformer-based models. Combining extractive and abstractive techniques to make hybrid models for summarization. Abstractive summarization can suffer from repetition and semantic irrelevance, causing grammatical errors and a poor reflection of the main idea of the source text. The disadvantage of ATS models is that this sequence of keywords is challenging to meet the requirement of syntactic structure. The other fundamental problem of mainstream ATS models is rare or OOV words.

1.1 Scope of this survey

For the last decade, the field of text summarization has been a trending area of research, leading to many summarization techniques surveys being published. Our objective is different; we propose a survey of text summarization in Indian Language which remain in their initial stage.

Item-based solutions generate the summary out of the top-rated sentences by employing the notion of information item (the smallest unit of textual information such as subject, verb, and object triplets) (Genest and Lapalme, 2011). On the other hand, Semantic-based approaches produce the desired summaries by extracting ontological and syntactical relations in text using semantic graphs, mainly by reducing the chart or locating its key concepts (Moawad and Aref, 2012). The top-ranked structure of text forms the summary (i.e., verbs, subjects, and objects) in the predicate argument-based approaches, which merge these respective structures (Zhang et al., 2016, Alshaina et al., 2017). Nevertheless, existing methods cannot achieve comparable performance to deep learning approaches. Thus, deep learning architectures have been widely adopted in abstractive Text summarization, and they have since become state-of-the-art (Gupta and Gupta, 2019) In abstractive text summarization, seq2seq models of encoder-decoder architectures are widely used for deep learning systems. It works along with attention mechanisms. Recurrent neural networks (RNNs) on extended short-term memory networks and gated recurrent units are primarily used in this system (Nallapati et al., Chopra et al., Chen et al., 2016; See et al., 2017, Song et al., 2018, Gupta and Gupta, 2019). In these cases, the sequence of words is the input of an encoder converted into a vector, and the attention mechanism at the decoder, which focuses on specific words, determines the output by emitting the next word of the summary based on the previous ones. (Bahdanau et al., 2014)

The above methodology is further extended in the neural attention-based model, which is trained end-to-end on a large amount of data that produces abstractive summaries. Similarly, Nallapati et al. (2016) and See et al.(2017) train encoder-decoder models with attention mechanisms incorporating a pointer generator network to face the problem of out-of-vocabulary (OOV) words. Further, the inclusion of a coverage mechanism and proposing a model of a convolution avoid repetition of the exact words in the summary(See et al. (2017), Lin et al. (2018) respectively). Finally, a deep LSTM-CNN (convolution neural network) framework generates summaries via the extraction of phrases from source sentences (Song et al. (2018)).

The deep learning-based model gives better performance because of the semantics concerned with it. [34] With the coming out of deep learning as a feasible alternative for many NLP tasks, scholars have started considering this framework attractive. Recurrent neural network-based sequence-to-sequence attention models have proven effective in abstractive text summarization.[34]

1.2 Motivation

If we consider research for summarization in Indian Language mostly work done of extractive, now researcher move towards the abstractive summarization. Various generic single document or multi- document abstractive-based summarization techniques are present. Majority are in English or other languages but was found to have very little work for the Indian Regional language. It has thus motivated us to check, analyze and compare the existing techniques so that abstractive summarization techniques for Indian Language can be proposed. Deep learning techniques are language agnostic and hence can overcome various shortcomings. In this section, we discuss some works on abstractive text summarization. [37,27]

2. SURVEY METHODOLOGY

This section exhibits the methodology employed to conduct this survey.

2.1 Survey plan

Pre-planning is required before conducting an exhaustive survey (Kitchenham et al., 2009). Therefore, We began with survey planning, identifying various research questions, determining relevant data sources, selecting keywords for research material searches, and deciding on the criteria for material inclusion and exclusion. We reviewed existing literature to carry out this systematic survey. Prior to conducting the proposed survey, relevant studies, publications, and articles were gathered and thoroughly assessed for quality.

2.2 Research questions

Some of the identified research questions used for the systematic survey, along with their objectives are listed in Table 2.

Table 2 Identified research questions and their aims for systematic survey

Q. No.	Identified research questions	Objective
RQ-1	How effective are existing summarization models for Indian languages	Its goal is to evaluate the performance of existing extractive and abstractive summarization techniques
RQ-2	How to create a robust dataset for text summarization in Indian languages?	Its goal is to design a diverse and large-scale corpus covering multiple Indian languages for training and evaluation of summarization models
RQ-3	What are the application domains, methodologies used,	Its goal is to present existing method and information on research opportunities in the field.
RQ-4	What are the challenges in text preprocessing for Indian languages	Its goal is to identify and address challenges like morphological richness, spelling variations, and lack of standardized grammar.
RQ-5	What evaluation metrics are suitable for Indian language summarization	To identify and adapt evaluation metrics that account for the linguistic and cultural nuances of Indian languages.
RQ-6	What role does semantic representation play in summarizing Indian languages	To investigate the effectiveness of semantic representations, such as word embeddings and contextual embeddings, in improving summarization quality.

2.3 Source selection

To ensure the credibility of this systematic survey, we selected only reputable and reliable data sources. We explored various platforms, including IEEE Xplore, ACM Digital Library, ScienceDirect (Elsevier),

Scopus, and Google Scholar, to compile a comprehensive bibliography of research papers on Summarization Techniques.

2.4 Inclusion and exclusion criteria

We select inclusion and exclusion criteria proposed by Gupta et al. (2020) to include and exclude articles for our survey. Figure 1 shows that the filtration procedure is divided into multiple phases. Various phases of the inclusion criteria were established based on the title, abstract, conclusion, full text, and identified challenges. Articles that did not meet these criteria were excluded, ensuring that only relevant studies were selected. Ultimately, only articles with significant citations were considered.

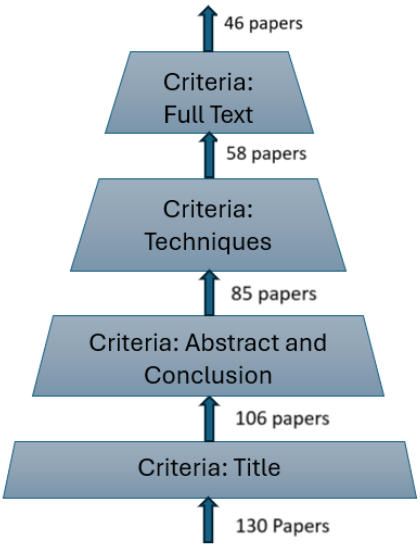

Figure: 1 Inclusion and exclusion criteria

3. THE NEED FOR TEXT SUMMARIZATION IN INDIAN LANGUAGES

Text summarization in Indian languages is essential for various reasons, includes

- *Information Overload Management*
  With the vast amount of news, government policies, legal documents, and research articles available in multiple Indian languages, summarization helps users grasp key insights quickly.
- *Enhancing Accessibility*
  Many people prefer concise information in their native languages, making summarization crucial for easy comprehension.
- *Bridging the Digital Divide*
  A significant portion of India's population consumes content in regional languages, but a lot of digital content is still in English. Summarization helps in making content accessible and relevant.
- *Improving Education & E-learning*
  Summarized study materials in regional languages make learning more efficient for students who may struggle with lengthy textbooks or research papers.
- *Legal & Governmental Communication*
  Government policies, court judgments, and legal documents are often lengthy. Summarization in Indian languages can help citizens understand important updates without going through complex legal jargon.
- *News & Media Consumption*
  With multiple news sources in different languages, automated summarization ensures that people receive the most relevant information in a short, understandable format.

- *Customer Support & Business Applications*
  Businesses operating in multilingual regions benefit from summarizing customer queries and responses, improving efficiency in communication.

4. THE STAGES OF TEXT SUMMARIZATION

The stages of text summarization typically follow a structured process to transform input text into a concise and meaningful summary. Figure 2 shows a breakdown of the key stages for text summarization.

4.1 Preprocessing

Preprocessing is essential in text summarization because Indian languages have unique linguistic characteristics such as complex morphology, diverse scripts, and agglutinative word formation. Without preprocessing, raw text can introduce noise, reducing the accuracy of summarization models. So before summarizing, the text needs to be cleaned. There are various pre-processing tasks, as shown in Table 3. Some approaches did not comment on the pre-processing tasks, while others used some of the pre-processing steps.

Table : 3 Key Preprocessing Tasks for Indian Language Text Summarization

Task	**Purpose**
Tokenization	Splits text into words/sentences for analysis.
Stopword Removal	Removes frequently used words (e.g., "का", "के", "और") that do not add meaning to the summary.
Stemming/Lemmatization	Converts words to their root forms to reduce variations. Example: "खेलना", "खेला" → "खेल".
Normalization	Handles spelling variations, Unicode normalization, and diacritic removal to maintain consistency
Named Entity Recognition (NER)	Identifies and preserves important names, places, and organizations.
Sentence Segmentation	Breaks text into meaningful sentences to improve summarization accuracy.
POS Tagging	Helps in understanding the role of words to improve summary quality.
Transliteration	Converts words between scripts, useful for code-mixed text (e.g., Hindi written in Latin script)

These preprocessing steps enhance the quality of input data, leading to more accurate and coherent summaries. Table 4 lists the available libraries to perform a pre-processing task for Indian language.

Table : 4 Available libraries to perform a pre-processing task for Indian language

Task	**Library**
Tokenization	IndicNLP, iNLTK, spaCy, Stanza
Stopword Removal	IndicNLP, NLTK
Stemming/Lemmatization	iNLTK, IndicNLP
Normalization	IndicNLP, NLTK
Transliteration	IndicTrans, IndicNLP, AI4Bharat
Sentence Splitting	IndicNLP, NLTK

4.2 Text Representation

Before summarization, text must be represented in a structured way that models can process. Different representation techniques are used based on which approach is used (extractive or abstractive summarization). To convert text into a structured form we need to convert text to numeric that is vectorization. For this some of the methods like TF-IDF, Word2Vec, BERT embedding are generally used. In extractive summarization sentence embeddings like BERT, IndicBERT, mBERT, LaBSE, etc. are used to capture contextual meaning by encoding whole sentences which works well for Indian languages. While in abstractive summarization Seq2Seq architecture with encoder-decoder model is commonly adopted in deep learning method where transformer-based representations like mT5, IndicBART, mBART-50 are commonly used[41]. This structured process helps in generating high-quality summaries for Indian languages while addressing linguistic complexities.

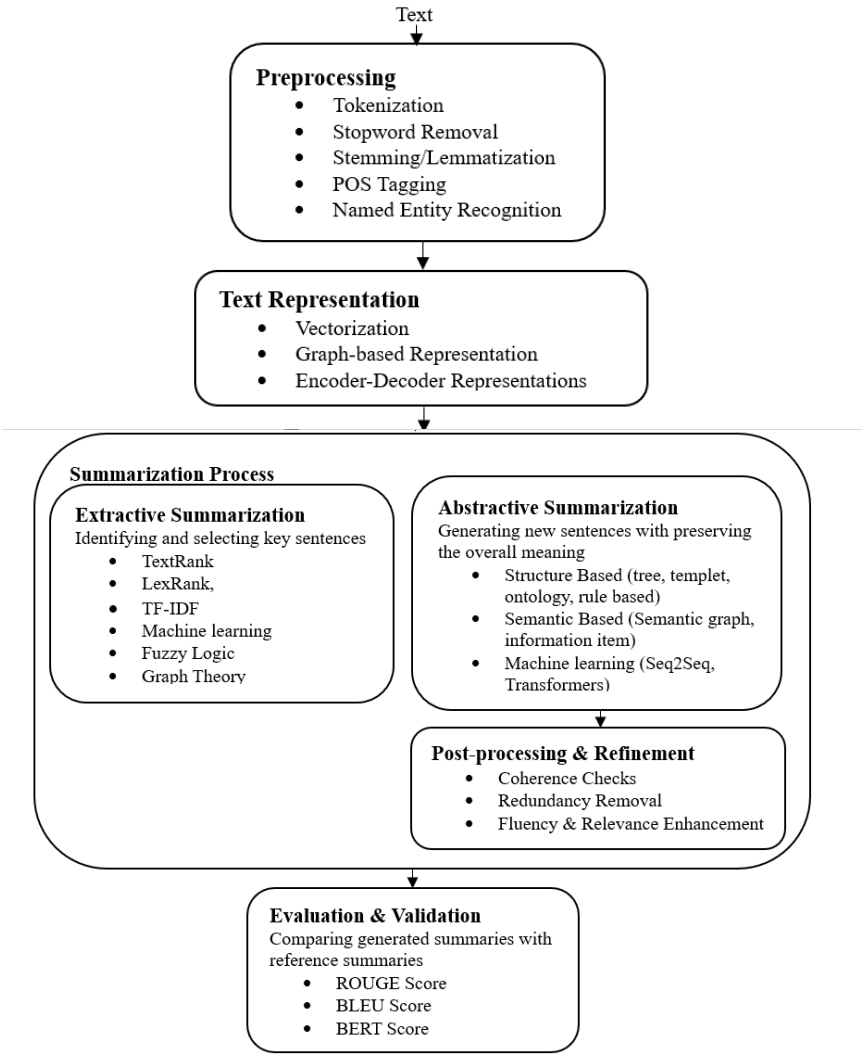

Figure: 2 The stages of the Text Summarization

4.3 Summarization Process

From main two approaches, extractive summarization uses sentence scoring and select high ranked sentences as a summary. TF-IDF, graph-based algorithms like TextRank, PageRank and machine learning models like Random Forest, SVM are used for sentence scoring. And abstractive summarization generates a new summary in a human-like manner, rather than copying sentences by using transformer-based models (mT5, IndicBART) to encode an input text, and model generates a new summary sentence-by-sentence. To ensure the grammar corrections, redundancy and fluency or relevance development some postprocessing task need to perform in abstractive summarization.

4.4 Evaluation & Validation

Evaluating a text summarization model in Indian languages requires both quantitative (automatic metrics) and qualitative (human evaluation) approaches. Evaluating summarization quality is crucial to ensure that the summaries generated are accurate, concise, and meaningful. For that multiple evaluation metrics are used. Indian languages present additional challenges such as morphological complexity, script diversity, and code-mixing. Below are key methods to evaluate summarization models for Indian languages. To evaluate the system generated summary we compare generated summaries with reference summaries based on below measure

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score that measures recall and precision against reference summaries. It works well for extractive summarization but doesn't consider semantic meaning.

BLEU (Bilingual Evaluation Understudy) Score that evaluates text fluency and accuracy based on word sequence similarity by measuring n-gram precision. It works well when reference summaries are short but not ideal for abstractive summaries in Indian languages

METEOR (Metric for Evaluation of Translation with Explicit Ordering) that considers semantic similarity and stemming which is more useful for Indian languages. But it heavily depends on the availability of linguistic resources.

BERTScore uses contextual word embeddings (BERT/IndicBERT) to compare semantic similarity. works better than ROUGE for abstractive summarization. But it is more computationally expensive.

Since automatic metrics don't always capture meaning accurately, human evaluation is necessary. Quality assessment-based evaluation like mean opinion score (MOS) in which human judges rate a summary in scales 1(low) to 5(high).

5. LITERATURE REVIEW ON INDIAN LANGUAGE

Bengali Text summarization:

M Islam et. al[15] proposed a hybrid extractive text summarizer, which generates a summary based on keyword scoring, sentiment analysis, and the interconnection of sentences. S Abujar et. al[16] proposed a heuristic approach in which scoring methods were implemented based on association and linguistic rules. R Sikder, et. al[17] include some basic steps of natural language processing along with statistical and mathematical techniques for grammatical analysis to design automatic extractive text summarization. As extractive text summarization on the Bengali language is proposed, researchers forward their work in abstractive.

Sheikh, et. al[2] presented an abstractive text summarization for Bengali text generation using LSTM-RNN. They discussed n-gram language modelling for text generation and created an RNN for the training model. Before executing a proposed model includes data collection, pre-processing, n-gram tokens sequence, and pad sequence. However, it quickly generated a fixed length and meaningful Bengali text. Md Ashrafu, et. al [3] proposed a sequence-to-sequence model to generate a summary. The model works with bi-directional RNNs with LSTM in the encoding layer and an attention model at the decoding layer. The main goal of the proposed model was to reduce the training loss of an abstractive summarizer and build a meaningful, fluent, and understandable summary. Prithwiraj, et. al[4] presented a seq2seq model based on Long Short-Term Memory (LSTM) network with an attention mechanism at encoder and decoder. The proposed model produces well-spoken and human-like generated sentences. They prepared a dataset of more than 19k articles and corresponding human-written summaries. Evaluation of model showed significant improvement with state-of-the-art. R R Chowdhury, et. al[5] proposed a graph-based unsupervised abstractive summarization system for Bengali text documents, which requires a pre-trained language model to train Bengali texts and a Part-Of-Speech (POS) tagger. They design a summarization tool to provide an extractive and abstractive summary of the Bengali document.

Malayalam Text summarization:

M Kondath, et. al [18] propose an extractive LDA-based topic modelling-based multi-document text summarization approach, which is employed to extract dominating topic terms from texts. Krishnaprasad P et. al [20] proposed the extractive summarization method of sentence extraction by ranking based on the word score in a single produces a generic summary for a given Malayalam document. Kanitha D K, et. al [19] proposed using a graph-based Malayalam text summarization. Now research domain in text summarization is going towards abstractive as extractive text summarization in the Malayalam language is already proposed.

K Kishore, et. al [6] present an attention-based abstractive text summarization approach that accepts a single document as input, processes the input by suitable semantic representation and then generates the final summary by using sentence framing techniques. The framework combines pre-processing, POS tagging, Karthav Karmam Kriya triplet extraction, and Karaka tree construction, merged via sentence aggregation rules. Finally, the sentences are developed according to grammar to generate the summary. S K Nambiar, et. al [7] proposed an abstractive summarization for Malayalam documents using a sequence-to-sequence model. The system uses the various compositional feature and inherent semantics as part of document- independent features. The model generates an understandable summary and reduces the training loss. Sunitha C, et. al [8] proposed abstractive summarization of Malayalam documents using the clause identification method. They had developed a POS tagger and a morphological analyzer for Malayalam words. The modified clause identification algorithm is used to identify clauses. Then semantic triples (subject, object, and predicate) analyze the clause semantically. And last, the score of each clause is calculated using feature extraction. R Kabeer et. al [9] developed two summarization methods that create generic text summaries for Malayalam documents. The extractive method uses standard statistical measures to rank the sentences. The second method, abstractive in nature, performs detailed semantic processing of the document to generate the summary.

***Kannada Text summarization*:**

Jayashree. R, et. al [21] K proposed document summarization in Kannada using keyword extraction by combining GSS (Galavotti, Sebastiani, Simi) coefficients, IDF methods, and TF for extracting keywords. A Swamy, Srinath S[22] developed the method for extractive Kannada text summarization using five sentence features which are TF-IDF, Sentence Frequency, Keywords feature, Sentence length, and Sentence position. . Jayashree, et. al [23] design the Categorized Text Document Summarization in the Kannada Language by Sentence Ranking. Abstractive summarization has been explored in the Kannada language as much work in extractive overview has been done.

V R Embar, et. al[10] proposed sArAmsha - a Kannada abstractive summarizer using a guided summarization approach. The sArAmsha system analyzes the document and performs POS tagging and stemming operations, using the abstraction schemes and Information Extraction (IE) rules to extract information and create a summary by forming sentences based on domain templates. Shilpa G V et. al [12] proposed a process that generates the abstractive summary for Kannada text documents by focusing on a unified presentation model based on Information Extraction (IE) rules and scheme-based Templates. The proposed method uses a custom-designed IE module categorization, rule-based, content selection, and sentence generation; they used TF/IDF rules for classification. Further lexical analysis reduces prolixity to robust IE rules. Templates for sentence generation make the summaries brief and information-intensive, generating a syntactically accurate, clean, and to-the-point summary. Work done by Geetha J K et. al [11] for text summarization is based on The Latent Semantic Analysis (LSA) that captures semantic relationships between the sentences as a human automatically. to generate the relevance to an original text document and non-redundancy in machine-generated summary. They use Singular Value Decomposition (SVD). It finds the mutually orthogonal dimensions of the Sentence.

Telugu Text summarization:

M. H khanam et. al [24] propose an extractive text summarization technique to summarize a Telugu document using a Frequency-based approach and K-means clustering. Sana Shashikanth and Sriram Sanghavi[25] show n frequency-based Text Summarization technique in Telugu and Foreign Languages and achieve a meaningful full result. K U Manjari [26] proposed the TextRank Algorithm for Extractive Summarization of Telugu Documents. Many algorithms and methods for extractive text summarization in the Telugu language are proposed, but very little work is done in abstractive text summarization.

Jagadish et. al [13] presented a paper of abstractive text summarization for Telugu, a south Indian regional language. This paper proposed an approach that includes Pre-processing, Summarizer, and Post-processing to summarize an input document. They conclude that Telugu has more complex morphological variations when compared with English. So, Different machine learning approaches are to be addressed to get more appropriate summaries for the given document. M Bharath, et. al [14] implemented an abstractive text summarization for the Telugu language using a deep learning model. They proposed an architecture based on encoder-decoder sequential models combined with an attention mechanism. They created the dataset manually and applied the model to generate a one-sentence summary and got good results measured qualitatively.

Other text summarization

A Raza et. al show the LSTM as encoder/decoder approach is a promising method for abstractive summary generation in Urdu[39] and produce summaries that are grammatically correct and semantically meaningful. S. Madria et al.(2019) give a comparative analysis and evaluate the effectiveness of different types of stemmers for the Gujarati language. Pinkeh et al. (2014) introduce the pre-processing phase of text summarization for Gujarati language. To construct summaries from Gujarati text. Shah and Patel use Textblob and Gensim for Gujarati text summarizer, This work is foundational for comprehending the linguistic and contextual intricacies involved in summarizing Gujarati. [45]. Patel examines the preprocessing phase for text summarization of Gujarati texts, emphasizing related issues and appropriate solutions.

Summary for abstractive summarization method on some Indian languages based on their advantage and disadvantage are brief in Table 5.

Table 5: Advantages and Disadvantages of summarization method in various Indian Language

Language	Method	Advantage	Disadvantage
Bengali	Sequence to Sequence RNN	An understandable, fluent, short, and meaningful summary that reduces the training loss	Summary with limited words
Malayalam	Sequence to Sequence approach	Increase efficiency and reduce the training loss.	Fix the length of the source and summary
Kannada	Template based	Good linguistic quality.	Normalization of Non-Standard Words needs to be handled.
Telugu	Deep Learning Neural Network.	Semantically relevant good results	Small data set.

6. ISSUES AND CHALLENGES

Text summarization in Indian languages is a rapidly evolving field with several challenges and opportunities. While extractive and abstractive methods have shown promising results, there is a need for more diverse datasets and language-specific models. Some of the common issues that need to faced by almost all Indian language summarizer, that are

- Low-resource language support (limited annotated corpora for Indian languages like Northeast Indian languages)
- Code-mixed text handling (Many users mix English with regional languages like Hinglish, Tanglish)
- Morphological Complexity (Indian languages have rich inflectional and complex linguistic structures.)
- Domain-specific summarization (like legal, medical)
- Dialectal Variations: Different states have variations of the same language.

Future research should focus on improving and enhancing summarization accuracy by dealing with these problems. There are some basic solution shown in table 6 suggested to cope up with these problems.

Table : 6 Challenges in Evaluating Indian Language Summarization

Challenge	**Solution**
Lack of large datasets (Limited annotated data for training deep learning models.	Creating benchmark datasets like XL-Sum (BBC), Indic-Sum
Morphological complexity (making tokenization and stemming harder)	Using embeddings (IndicBERT, mBERT) instead of word overlap
Code-mixed text (Hinglish, Tanglish, etc.)	Custom tokenization and transliteration
Lack of standard evaluation tools	Adapting existing tools for Indian languages

7. APPLICATION DOMAIN

Text summarization in Indian languages has various applications across multiple domains due to the diversity of languages spoken in India. Here are some key application domains:

1) *News and Journalism:* Automatic summarization of news articles in Hindi, Tamil, Bengali, and other regional languages. Generation of news digests for regional audiences. Personalized news recommendations based on user preferences.
2) *Government and Policy Documents*: Summarization of government policies, circulars, and legal documents in multiple Indian languages. Easy-to-read summaries for citizens to understand regulations and benefits.
3) *Education & E-Learning*: Condensing long academic texts and research papers into concise summaries. Automated generation of summaries for regional-language study materials. Helping students grasp key points in their native language.
4) *Healthcare & Medical Information*: Summarizing medical reports and health articles in regional languages for better accessibility. Providing short, easy-to-understand summaries of prescriptions and medical guidelines.
5) *Legal Domain*: Summarization of court judgments and legal case reports in languages like Hindi, Marathi, and Tamil. Helping lawyers and citizens access legal information quickly.
6) *Customer Support & Chatbots*: Summarizing customer queries and responses for efficient support in multilingual settings. Providing concise answers to frequently asked questions in Indian languages.
7) *Social Media & User-Generated Content:* Summarizing trending topics and discussions on social media platforms in regional languages. Condensing long opinion pieces or blog posts for better readability.
8) *Business & Market Analysis:* Summarizing financial reports and market trends in local languages. Providing insights from consumer reviews in different Indian languages.
9) *E-Governance & Public Services:* Summarizing government schemes, public service announcements, and citizen feedback. Enhancing accessibility for rural and non-English speaking populations.

10) *Agriculture & Rural Development*: Summarizing agricultural guidelines, weather updates, and market prices for farmers. Delivering key insights from research papers in local languages.

8. TOOLS AND DATASET FOR SUMMARIZATION IN INDIAN LANGUAGE

Indian language summarization involves processing text in multiple languages, each with unique linguistic structures, grammar, and vocabulary. There are several tools and models available for text summarization in Indian languages. Here are some of the best options listed in table 7.

Table : 7 Tools for Indian Language Summarization

Library	**Functions**
IndicNLP (AI4Bharat)	Supports multiple Indian languages for NLP tasks
iNLTK (Indian NLP Toolkit)	Provides tokenization, summarization, and other NLP tasks for various Indian languages
Stanford NLP	Offers dependency parsing, named entity recognition (NER), and summarization
spaCy with IndicNLP	Can be fine-tuned for Indian languages
IndicBERT, mBERT, and XLM-R	For language understanding and summarization[43]
Sumy (Python Library)	Useful for NLP tasks including summarization
Hugging Face Transformers	Provides pre-trained models for summarization
mBART (Multilingual BART)	Works well for abstractive summarization
FastText	Supports Indian languages for text processing and embedding
IndicTrans	Neural machine translation model for Indian languages
Google's MuRIL	Trained on various Indian languages for NLP tasks.

There are several datasets available for summarization in Indian language

- IndicCorp – Large-scale corpus for Indian languages.
- AI4Bharat Datasets – Provides labeled data for NLP tasks.
- WikiLingua – Multilingual Wikipedia-based summarization dataset.
- NewsOnAir (AIR) – News summaries in Indian languages.
- IISc & IIT Datasets – Research datasets for NLP in India.

9. RESEARCH OPPORTUNITIES

Text summarization in Indian languages presents several research opportunities due to the linguistic diversity, rich morphology, and scarcity of high-quality annotated datasets. Though several approaches have been proposed, there are still several substantial associated research areas that are yet to be explored as mentioned below.

1. Multilingual and Low-Resource Summarization: Many Indian languages, such as Assamese, Odia, and Manipuri, have limited digital resources. Research can focus on transfer learning, zero-shot learning, and few-shot learning for summarization in low-resource Indian languages. Research on leveraging models like mBART, IndicBERT, and MuRIL for summarization across multiple Indian languages can further be enhanced.
2. Abstractive Summarization Challenges: Indian languages often have free word order and complex morphological structures, making abstractive summarization difficult. Models trained on English perform poorly due to lack of large-scale summarization datasets in Indian languages. Developing transformer-based approaches specifically for Indian languages using IndicNLP embeddings is a further research direction in abstractive summarization.

3. Code-Mixed Summarization: Many Indian users use code-mixed text (Hindi-English, Tamil-English, etc.) in social media, news, and conversations. Research is needed in handling code-switching and transliteration in summarization tasks. For this developing synthetic code-mixed summarization datasets is a challenges task.
4. Domain-Specific Summarization: Legal, medical, and financial documents in Indian languages need customized summarization approaches. It can further be enhanced by Creating domain-adapted summarization models for Indian languages using BERT variants trained on legal or medical text.
5. Dataset Creation and Benchmarking: Lack of high-quality, large-scale summarization datasets is a major challenge. Initiatives like IndicNLP and AI4Bharat can be extended to create summarization datasets. To Build the corpus use crowdsourcing and automatic data mining technique.
6. Evaluation Metrics for Indian Languages: Traditional metrics like ROUGE may not be effective due to rich morphology and flexible word order in Indian languages. Research can focus on new evaluation metrics that consider semantic similarity and grammatical correctness.

10. CONCLUSION

In recent years, as there is a huge availability of data on the web the importance of the text summarization process has increased. Extractive and abstractive methods are the two types of text summarization. This paper tries to review the strategy of abstractive text summarization. The concept of abstractive text summarization understands the document and rephrases the original text to new phrases to generate a summary close to the human. There is a limitation work carried out in Indian language processing. This paper shows complete flow of text summarization for Indian Language. It gives the opportunities challenges and future prospect of Indian language summarization. Based on this critical survey it justified that deep learning approaches are adopted for abstractive text summarization. And we conclude that Encoder-Decoder architecture is mainly used to solve the Seq2Seq problems of text summarization. Some approaches applied LSTM to solve gradient vanishing problem. For Indian language combination of RNN and attention mechanism are the most used deep learning techniques, which gives a good performance compared to the conventional methods. Regional Language needs more linguistic processing. Even though there is no word structure, it requires handling inflection, which plays the most crucial role in summarization. There is no strong abstractive method for Indian regional Language because individual words of every sentence access domain ontology & word net. Thus, we suggested some tools for Indian Language Summarization.

References

[1] Kishore Kumar Mamidala, Suresh Kumar Sanampudi, "Text Summarization for Indian Language: A Survey", International Journal of Advanced Research in Engineering and Technology (IJARET) 2021 Volume 12, Issue 1, pp. 530-538

[2] Sheikh, Abu Kaisar, Sanzidul Islam, Fahad and Syed, "A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN", Seventh International Conference on Innovations in Computer Science and Engineering (ICICSE), Springer - 2019.

[3] Md Ashrafu, Sheikh, Abu Kaisar, Fahad Faisal, Syed Akhter Hossain, "Bengali abstractive text summarization using sequence to sequence RNNs", 10th International Conference on computing communication and network Technologies(ICCCNT), IEEE- 2019.

[4] Prithwiraj, Avi, Md Saifu and Marium-E- Jannat, "Bengali Abstractive News Summarization(BANS): A Neural Attention Approach", arXiv:2012.01747[cs.CL], 2020

[5] R R Chowdhury, M T Nayeem, Tahsin and Md. Saifu,"Unsupervised Abstractive Summarization of Bengali Text Documents", 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2612– 2619 April 19 - 23, 2021.

[6] Kavya Kishore, G N Gopal and Neethu, "Document summarization in Malayalam with Sentence Framing", International Conference on Information Science (ICIS) - 2016.

[7] Sindhya K Nambiar, David Peter S, Sumam Mary Idicula, "Abstractive Summarization of Malayalam Document using Sequence to Sequence Model", 7th International Conference on Advanced Computing & Communication Systems (ICACCS) , IEEE 2021.

[8] Sunitha C, A Jaya, Amal Ganesh, "Automatic summarization of Malayalam documents using clause identification method", International Journal of Electrical and Computer Engineering (IJECE), 2019, Vol.9, ISSN: 2088-8708, pp. 4929-4938.

[9] Rajina Kabeer and Sumam Mary Idicula, "Text Summarization for Malayalam Document - An Experience", International Conference on Data Science & Engineering (ICDSE) - 2014.

[10] Varsha R Embar, Surabhi R Deshpande, Vaishnavi A K, Vishakha Jain, Jagadish S Kallimani, "sArAmsha - a Kannada Abstractive Summarizer", International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2013, PP : 540-544

[11] Geetha J K and Geetha J K ,"Kannada Text summarization using Latent Semantic Analysis ", International Conference on Advances in Computing, Communications and Informatics (ICACCl) - 2015, PP: 1508-1512

[12] Shilpa G V and Shashi Kumar D R, "Abs- Sum-Kan: An Abstractive Text Summarization Technique for an Indian Regional Language by Induction of Tagging Rules", International Journal of Recent Technology and Engineering (IJRTE) - 2019, ISSN: 2277-3878, Volume-8, Issue- 253.

[13] Jagadish, Srinivasa and Eswara Reddy, "Information Extraction by an Abstractive Text Summarization for an Indian Regional Language", 7th International Conference on Natural Language Processing and Knowledge Engineering IEEE 2011, pp: 319-322

[14] Mohan Bharath, Aravindh Gowtham, Akhil M, "Neural Abstractive Text Summarizer for Telugu Language", arXiv:2101.07120 [cs.CL], 2021

[15] Mahimul Islam, Fariha Nuzhat Majumdar, Asadullahhil Galib, Md Moinul Hoque, " Hybrid Text Summarizer for Bangla Document", International Journal of Computer Vision and Signal Processing, PP: 27-38, 2020

[16] Sheikh Abujar, Mahmudul Hasan, M.S.I Shahin, Syed Akhter Hossain , " A Heuristic Approach of Text Summarization for Bengali Documentation", 8th International Conference on computing communication and network Technologies (ICCCNT), IEEE- 2017.

[17] Ratul Sikder, Md. Monowar Hossain, F.M. Rahat Hasan Robi, "Automatic Text Summarization For Bengali Language Including Grammatical Analysis", International Journal of Science & Technology Research, Volume 8, ISSUE 06, ISSN 2277-8616, 2019

[18] Manju Kondath, David Peter Suseelan, and Sumam Mary Idicula, "Extractive summarization of Malayalam documents using latent Dirichlet allocation: An experience", Journal of Intelligent Systems 2022; Issue : 31, PP: 393–406

[19] Kanitha D K, M N Mubarak, S.A. Shanavas, " Malayalam Text Summarization Using Graph Based Method ", International Journal of Computer Science and Information Technologies(IJCSIT), Vol.9 Issue 2,

ISSN:0975-9646, 2018, pp:40-44

[20] Krishnaprasad P, Sooryanarayanan A and Ajeesh Ramanujan, "Malayalam Text Summarization: An Extractive Approach ", International Conference on Next Generation Intelligent Systems (ICNGIS), IEEE-2016

[21] Arpitha Swamy, Srinath S, "Automated Kannada Text Summarization using Sentence Features", International Journal of Recent Technology and Engineering(IJRTE) ISSN:2277-3878, Volume-8 Issue-2, July 2019.

[22] Jayashree, Srikanta Murthy K and Basavaraj .S.Anami, "Categorized Text Document Summarization in the Kannada Language by Sentence Ranking",12th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE - 2012, pp: 774-781.

[23] Dr. M. Humera khanam and S. Sravani, "Text Summarization for Telugu Document", Journal of Computer Engineering (IOSR-JCE), ISSN: 2278-8727, Volume 18, Issue 6, 2016, PP 25-28.

[24] Sana Shashikanth, Sriram Sanghavi "Text Summarization Techniques Survey on Telugu and Foreign Languages ", nternational Journal of Research in Engineering, Science and Management (IJRESM) Volume-2, Issue-1, January- 2019 ISSN : 2581-5792, PP:211-213

[25] K Usha Manjari, "Extractive Summarization of Telugu Documents using TextRank Algorithm", 4th International Conference on IoT in Social, Mobile, Analytics and Cloud (I-SMAC) IEEE; ISBN: 978-1-7281-5464-0, PP: 678- 683.

[26] Sunitha C, Dr. A Jaha, Amal Ganesh, "A Study on Abstractive summarization Technique in Indian Language", 4th International Conference on recent trends in Computer Science & Engineering. ELSEVIER - 2016 PP: 25-31

[27] Rike Adeliaa, Suyanto Suyantoa, Untari Novia Wisesty, "Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit",4th International Conference on Computer Science and Computational Intelligence (ICCSCI), 2019, PP: 581-588.

[28] Jagadish S KALLIMANI, Srinivasa K G, Eswara REDDY B, "Information Extraction by an Abstractive Text Summarization for an Indian Regional Language", IEEE- 2011. PP: 319-322

[29] Vishal Gupta, "A Survey of Text Summarizers for Indian Languages and Comparison of their Performance", Journal of Emerging Technologies in Web Intelligence, VOL. 5, NO. 4, 201

[30] Nomi Baruah, Shikhar Kr. Sarma, Surajit Borkotokey, "Text Summarization in Indian Languages: A Critical Review", Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)-IEEE 2019

[31] Rahul, Shristi Rauniyar, Monika, "A Survey on Deep Learning based Various Methods Analysis of Text Summarization", Fifth International Conference on Inventive Computation Technologies (ICICT)-IEEE 2020 ISBN:978-1-7281-4685-0 PP: 113-116

[32] Surendrabikram Thapa, Sushruti Mishra, "Review of Text Summarization in Indian Regional Languages", 3rd International Conference on Computing Informatics an Network, Springer - 2021, PP: 23-32

[33] Rupal Bhargavaa, Gargi Sharmaa, Yashvardhan Sharma, "Deep Text Summarization using Generative

dversarial Networks in Indian Languages", International Conference on Computational Intelligence and Data Science (ICCIDS) ELSEVIER -2020 PP: 147-153.

[34] A. B. Al-Saleh and M. E. B. Menai, "Automatic Arabic text summarization: a survey", Artificial Intelligence Review, vol. 45, no. 2, pp. 203–234, 2016.

[35] D. Suleiman and A. A. Awajan, "Deep learning based extractive text summarization: approaches, datasets and evaluation measures", in Proceedings of the 2019 Sixth International Conference on Social Networks Analysis Management and Security (SNAMS), pp. 204–210, Granada, Spain, 2019.

[36] Sabina Yeasmin, Priyanka Tumpa, "Study of Abstractive Text Summarization Techniques", American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 Volume-6, Issue-8, pp : 253 -260, 2017.

[37] A. Khan, and N. Salim, A Review On Abstractive Summarization Methods, Journal of Theoretical and Applied Information Technology, 59(1), 64-71, Malaysia, 2014

[38] Ali Raza, Hadia Sultan Raja, Usman Maratib "Abstractive Summary Generation for the Urdu Language", arXiv:2305.16195v1 [cs.CL] 25 May 2023

[39] Abhinaba Bala, Ashok Urlana, Rahul Mishra, Parameswari Krishnamurthy, "Exploring News Summarization and Enrichment in a Highly Resource-Scarce Indian Language: A Case Study of Mizo", arXiv:2405.00717v1 [cs.CL] 25 Apr 2024

[40] Mrithula Kl, Aishwarya Krishnakumar, B. Bharathi, " Text summarization for Indian languages using pre-trained models", Conference on Forum for Information Retrieval Evaluation, Vol-3395 May 2023,

[41] Aparna Madhukar Mete, Manikrao Laxmanrao Dhore, "A Comprehensive Survey on Abstractive Text Summarization of Devanagari Script Based Hindi Language", International Journal of Intelligent Systems And Applications In Engineering, 2024, 12(3), pp 3604–3620

[42] Srisudha Garugu, D. Lalitha Bhaskari, "Enhancing Abstractive Text Summarization using Two-Staged Network for Telugu Language (EATS2N)", International Journal of Intelligent Systems and Applications in Engineering IJISAE, 2024, 12(19s), 686–695

[43] Dhaval Taunk, Vasudeva Varma, "Summarizing Indian Languages using Multilingual Transformers based Models", arXiv:2303.16657v1 [cs.CL] 29 Mar 2023

[44] Prakash Dhakal, Daya Sagar Baral, "Abstractive Summarization Of Low Resourced Nepali Language Using Multilingual Transformers" , arXiv:2409.19566v1 [cs.CL] 29 Sep 2024

[45] Malvi Shah, Kalyani Patel,"ગુજરાતી Gujarati Text Summarizer" International Research Journal of Engineering and Technology, 2019. https://www.irjet.net/archives/V6/i6/IRJET-V6I6230.pd

[46]Riddhi Kevat, Sheshang Degadwala, "A Comprehensive Review on Gujarati-Text Summarization Through Different Features", International Journal of Scientific Research in Computer Science Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 10, pp.301-306, September October-2023. DOI:10.32628/CSEIT2361051