International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Study of Marathi Text Summarization using Natural Language Processing

Vaishali S. Kapse[1], Dr. Sonal S. Deshmukh[2]

[1]Department of Master of Computer Applications, MGM University, Chhatrapati Sambhajinagar, India
Email: *vaishalikapse33[at]gmail.com*

[2] Department of Master of Computer Applications, MGM University, Chhatrapati Sambhajinagar, India
Email: *sonalsdeshmukh8[at]gmail.com*

Abstract: ***Text summarization is one of the most essential tasks in natural language processing (NLP), as it attempts to reduce big text bodies while retaining key information. Although much research has been conducted in this field for English texts, studies on other languages, such as Marathi, is absent. This study focuses on comparative study of various ways for extractive and abstractive automatic text summarizing and evaluates how well they capture the semantic core of Marathi text, as well as the complex nature of Marathi language structures and linguistic elements required for accurate summarization. It also analyses how summarization models function as they deal with the linguistic complexity and syntactic variation seen in Marathi texts, with a focus on creating clear yet useful summaries while accounting for context sensitisvity, word order, and morphological diversity. Each of these strategies has unique challenges that can be handled through using a specific variation of that strategy.***

Keywords: Text Summarization, Natural Language Processing (NLP), Marathi Language, Abstractive

1. Introduction

Every day, the amount of text available on the internet and in other libraries grows enormously. Because data is continually growing and contains noise and irrelevant material, accessing information has become expensive and time - consuming. Text summarization is a technique for summarizing data. There is no doubt that a manual text summary technique is effective in retaining the text's sense, but it takes time. Automatic text summarization is an additional approach (ATS). In ATS, computers may be programmed to generate information summaries using a variety of helpful techniques. As a result, text summarizing focuses on the most significant facts while keeping the flow of the original text, resulting in a brief and accurate overview. The primary purpose of text summary is to convey the content of a document in fewer words and sentences.

There are two types of summaries: extractive and abstractive. In extractive summaries, relevant sentences are extracted from the original text and put in the correct order. Relevant sentences are identified from input text using statistical and language - dependent features. In contrast, abstractive text summaries are constructed using natural language comprehension. However, mother tongues are much more than just languages. The Marathi language is rated fifteenth in the world, with 83 million native speakers in India alone—not to add the 95 million speakers who speak Marathi as their mother tongue. [14] It is both the official and regional language of Maharashtra. Even up until that time, not much research has been done on Marathi text summary. Marathi has been chosen as the study language as a result. Marathi is written using the Devanagari script, which has one of the largest letter sets.

In moment's information age, multitudinous scientific papers are released throughout colorful fields of exploration. In the ultramodern world, people prefer to read composition summaries to be informed about current events. Despite advancements in NLP approaches, Automatic Text Summarization remains the most difficult task. Text summarization seeks to prize underpinning meaning from lengthy textbooks. Long text or other sources are condensed into crucial information to save time, plutocrat, and trouble. [9]

1.1 Problem Statement

Produce an excellent text summarization system for Marathi language documents that will automatically induce short and useful summaries from enormous quantities of text. The system should be suitable to directly induce summaries while retaining the main information and general meaning of the original content. Apply or elect applicable text summarization ways, including extractive and/ or abstractive styles, for the Marathi language. Pre - process the Marathi text data by removing noise, punctuation, and stop words before tokenizing it into rulings or words for posterior analysis. Train and fine - tune the text summary model with the pre - processed dataset, optimizing for the delicacy, consonance, and readability of generated summaries.

1.2 Natural Language Processing

Natural language processing, or NLP, is a machine literacy approach that allows computers to comprehend, manipulate, and interpret mortal language. Organizations are presently gathering a lot of text and voice data from a range of communication channels, similar as emails, textbooks, social media newsfeeds, audio, videotape, and more. They use natural language processing (NLP) software to automatically reuse this data, assess the sentiment or intent of the communication, and respond to mortal discussion in real time. It's clear that different languages are spoken in different regions of the world. Nevertheless, individualities that fall into a particular group or live in a particular area speak or use a particular mortal language. The diversity of

natural languages is the main motorist of societal advancement and confluence. The only real way to ameliorate society and change the social atmosphere is generally through natural language, as it used to be called. NLP is a distinct area of artificial intelligence and computer wisdom that focuses on the study of mortal languages in the ultramodern digital age. [14]

Figure 1: Overview of NLP

2. Motivation

The purpose of this study is to speed up knowledge about NLPs techniques by providing an overview of current research, specifically in Text Summarization. It also makes possible to develop new resources, techniques, datasets, and tools to satisfy the demands of the industrial and research sectors. With the development of NLPs, automatic text summarization became practical for sentiment analysis and regular text document summaries. Additionally, text summarization encourages a flexible method of conducting research in a number of disciplines, including psychology, natural language, cognitive science, and machine learning. These combined results, which were obtained from a variety of sources, served as the inspiration for this study.

3. Text Summarization

In NLP, text summarization is a delicate exploration task that yields a precious summary of any given input document. That is, in fact, the process of reducing a long text document to a shorter interpretation while maintaining the original meaning and overall sense. A summary is a shorter interpretation of an important textbook. The size of this summary text is constantly lower than that of the main text. The primary task is to prize or choose the most important information from the original text, also represent that information to produce a summary. [14]

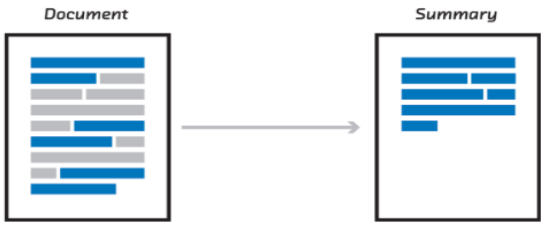

Figure 2: Text Summarization

There are three ways to text summarization extractive, abstractive, and hybrid. Each approach has multiple approaches and procedures. Each approach has both advantages and downsides. Figure 1 provides an overview of the ways and specific methodologies used. There are three ways to text summarization extractive, abstractive, and Hybrid. Each approach has multiple approaches and procedures. Each approach has both advantages and downsides. A brief summary of the methodologies, together with. Figure 1. Illustrates some specific ways.

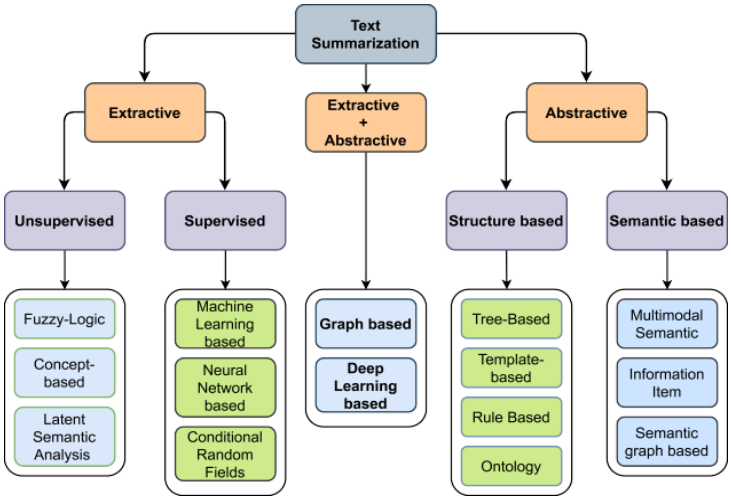

Figure 3: Classification of Text Summarization Techniques

3.1 Extractive Text Summarization

The extractive text summarization approach seeks to find words and sentences in a text and use them efficiently to generate a summary. This requires choosing sentences from the original document depending on its importance. These key sentences are then utilized to reproduce the text's core elements word for word, yielding a subset of the original document's phrases. [27]

Extractive text summarization employs two types of machine learning approaches: supervised and unsupervised machine learning, as illustrated in Figure. The next section provides a quick summary of various sub classes.

3.1.1 Supervised learning Methods

In supervised learning approaches, the initial stage is to learn how to classify documents by training to recognize summarized and non - summarized texts. These approaches'

machine learning and neural network algorithms require a categorized dataset for training, which includes both summarized and non - summarized texts with lables.

3.1.2 Unsupervised learning Methods

It allows for summarization without human assistance, such as picking initial sentences from a document. These solutions just require complex algorithms, such as graph - based, concept - based, fuzzy logic, and latent semantics, to accept user input and work automatically. These approaches are useful for handling large amounts of data. [21]

3.2 Abstractive Text Summarization

Abstractive text summarizing is the evolution and automation of the traditional approach of text summarization. The abstractive process involves paraphrasing important sections and main concepts from a literary piece. There are two sorts of abstractive summarization approaches: structure - based and semantic - based. A quick description of these two types based on natural language processing is provided below:

3.2.1 Structure - based Methods:

It filters crucial info from documents using abstract or cognitive algorithms. The algorithms for tree - based, template - based, and rule - based ontologies are the most widely used.

3.2.2 Semantic - based Methods:

This approach refines sentences by applying NLP to the entire manuscript. Some strategies can be used to readily find the noun and verb phrases. These methods include the multimodal semantic method (MSM), semantic graph - based method (SGM), information item - based method (IIM), and semantic text representation model (STRM). [27]

3.3 Abstractive + Extractive (Hybrid):

3.3.1 Graph based:

Both extractive and abstractive text summarization can be accomplished with the graph - based approach. Using a graph, this unsupervised learning technique assigns a rating to the necessary sentences or terms. It use LexRank, PageRank, TextRank & many more techniques

3.3.2 Deep Learning:

Information - driven ATS can be made more effective, accessible, and user - friendly with the use of deep learning models. Since these models aim to replicate human brain functions, they hold great promise for ATS. Since deep neural networks' architecture complements the complex structure of the language, they are frequently used to solve NLP problems

3.4 Challenges in the Summarization Task

After reviewing the literature, it becomes clear that researchers deal with numerous challenges while trying to fix NLP problems such as text summarization. Some significant challenges are highlighted below
.

1) **Lack of Standardization:** The Marathi language employs Devnagari scripts to write text, although each word is represented differently in its spellings or aksharas. It primarily depends on how the language is utilized by the speaker with its genuine intention and tone of speaking, such as whether it takes extended pronunciation or short pronunciation. [14]
2) **Phrases:** A collection of practical Marathi phrases makes the language unique, rich, and easy to learn. Because a phrase frequently uses certain words in the sentence, they have unique looks and their purposes separate from the terms' true meanings.
3) **Post positions:** Many words in the language have multiple forms, with some suffixes added to the fundamental nouns.
4) **Absence of capitalization:** By default, grammatical rules and specifications, the Marathi language does not have a concept of capitalization for writing nouns. It makes it difficult to find the key terms.
5) **Complicated Morphology:** Marathi has grammatical complexity since it is a free - ordered language with complex structures and syntax variations
6) **Lack of Clarity:** The same term frequently conveys a distinct meaning in that context, making language more difficult to understand.
7) **Code Mixed Data:** With original source language other language words or foreign words are used for communication which creates complexity for processing the standard language text. This code mix data is the additional foreign language words rather than Marathi. E. g. Marathi and Hindi, Marathi and English.

4. Literature Survey

We have looked into the surveys that are currently available in the text summarization area, and some of them are provided to demonstrate the importance of this work. The majority of surveys addressed earlier approaches and summarization research.

Sheetal Ajaykumar Takale (2023) provided a summary of the different methods that have been suggested for summarizing legal documents. Supervised, extractive, abstractive, AI - based knowledge representation, large language model based or pre - trained model - based approaches are the different categories into which approaches fall. Additionally, it has been noted that abstractive models consistently perform better than extractive. [29]

Mukesh Kumar Rohil et. al (2022) this paper reviews some recent discussions about the advantages and disadvantages of automated text summarization in the biomedical and healthcare domains. Along with providing various forms of support for researchers, their work additionally provided instruction on some implicit uses of text summarization in the biomedical and healthcare fields, with the goal of making medical documents and health records as readable as possible. [23]

Chetana Varagantham et. al (2022) offered a framework for condensing the vast amount of data, and their suggested approach relies on extracting highlights from the internet

using both semantic and morphological data. [30] Aakash Srivastava et. al (2022) they studied that the Search engines such as Google and Yahoo were developed in order to retrieve information from databases. The actual results have not been attained since the volume of electronic information is increasing daily. Automated summarization is therefore much looked after. This study was carried out in a single document. The frequency - based technique to text summarization is the main topic of this report. [27]

M. F. Mrida et. al (2021) provides a systematic survey of the broad ATS domain in different phases: performance measurement matrices, dataset inspections, important text summarization algorithms, feature extraction architectures, and fundamental theories with prior research backgrounds and challenges of current architectures and challenges. [21]
Ishitya Awasti et. al (2021) focused on reviews a number of research papers on learning methods, including supervised, unsupervised, and reinforcement learning, as well as abstractive, extractive, and hybrid techniques. They also examine extractive and abstract approaches to text summarization & techniques that result in a summary that is more focused and less repeated. [23]

Yogesh Kumar et. al (2021) this study provides a comprehensive summary of the research conducted on automatic text summarizing across multiple languages using different text summarization techniques., this study educates and supports beginner scientists in this field. [19]

Rahul C Kore et. al (2020) offered the fundamentals of the text rank algorithm, which determines the order of consequence for each sentence in a document by using the runner rank algorithm as a base. Next, the similar vector representation of each word was figured, and the vector representation of each judgment was attained by calculating a normal of all these vectors. [18]

Sheng - Luan Hou et. al (2020) has given the overview of important DL - based approaches that have been put forth in the text Summarization tasks and to trace their development. Firstly, provide an overview of DL and ATS. Following that, a summary of methods for multi - document summarization is provided. [13]

Nikhil S. Shirwandkar et. al (2018) they worked on a system in which fuzzy logic & RBM are used as an unsupervised learning algorithm to increase the summary's accuracy. This approach yields an 88% F measure, 80% recall, and 88% precision on average. When the suggested method is used instead of the current RBM method, better evaluation parameters are obtained [26].

Sumya Akter et. al (2017) proposed a technique for text summarization presented in this paper extracts key phrases from one or more Bengali documents. The K - means clustering algorithm has been used to generate the final summary for either one or more documents. [1]

Ambedkar Kanapala, et. al (2017) they try to survey various approaches to text summarizing that have been developed recently. They also discuss the salient features of several summarizing strategies. A few software applications used in legal text summarizing are briefly covered. At last, offer some suggestions for further investigation. [2]

Table 1: Literature survey of existing summarization systems in Marathi Language

Author	Dataset	Techniques / Methods	Results	Limitations / Future Scope
Mayank Mahajan et al. (2024) [20]	Marathi News	Text Rank Algorithm	Good summary accuracy	Improve accuracy further
Sunil D. Kale et al. (2023) [16]	Marathi Text Documents	TF - IDF	92.76% accuracy	Add advanced web crawler
Virat V. Giri et al. (2023) [10]	Marathi news on politics, economics, social topics	SVD & Fuzzy Logic	SVD better for multi - doc; Fuzzy for single - doc	SVD lacks semantics, fuzzy lacks context
Kirti P. Kakde et al. (2023) [15]	IndicNLP Marathi News	LexRank Algorithm	78% precision	Explore deep learning techniques
Utkarsh Hajare et al. (2023) [11]	Marathi Text Documents	Text Rank	Rouge4: Sim=0.7; Pos=0.8	Test more parameters and algorithms
Vaishali P. Kadam et al. (2022) [14]	Online Marathi Stories	Summarization Technique	44.48% compression accuracy	Improve for abstract summaries
Deepali Kadam (2021) [12]	Extractive Summarization Survey	Feature Extraction Survey	Feature extraction development needed	Still under development
Manasi Chouk et al. (2021) [7]	Marathi Text	Various Techniques	Good accuracy	Add features, try abstractive methods
Shruti Bhoir et al. (2021) [4]	Marathi Text Documents	Deep Learning	Low accuracy	Better feature selection needed
Apurva Dhawale et al. (2020) [8]	GitHub: 1135 Marathi articles	TextRank	Ratio between 0 and 1	Accuracy optimization needed
Anishka Chaudhari et al. (2019) [6]	1000 Marathi News Articles	Recurrent Neural Network (RNN)	Precision=0.319; Recall=0.342; Accuracy=0.32	LSTM could improve performance
V. V. Sarwadnya et al. (2018) [24]	634 Marathi News Articles	Feature Extraction, Scoring, Graphs	Rouge4: Sim=0.77; Pos=0.84	Add scoring & semantic ranking
Shubham Bhosale et al. (2018) [5]	Marathi E - News	Keyword Extraction	Efficient accuracy	Accuracy affected by text length
Yogeshwari V. Rathod (2018) [22]	Marathi News Articles	Unsupervised Keyword & Sentence Extraction	Sim=0.70; Pos=0.62	Effective summaries obtained

* Keys:
- Sim = Similarity - based score
- Pos = Position - based score
- TF - IDF = Term Frequency - Inverse Document Frequency
- SVD = Singular Value Decomposition
- LSTM=Long Short - Term Memory

In general, conventional algorithms such as TextRank and LexRank have produced accuracy and precision that range from moderate to good. For instance, when it came to Marathi news summaries, TextRank had a high accuracy rate. Deeper semantic understanding is frequently missing from methods that rely on simple statistical and rule - based models, like Term Frequency. Recurrent neural networks (RNNs) are one of the more sophisticated techniques that have been tested in some studies. Nevertheless, the outcomes of these methods have been variable; for example, RNNs performed less accurately, and fuzzy logic had difficulty with non - linear data in Marathi text. Following figure 4 shows various techniques and results for Marathi text summarization.

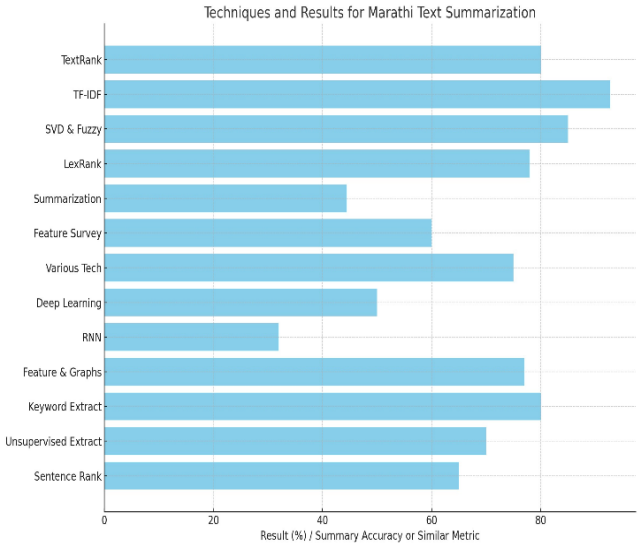

Figure 4: Techniques & results for Marathi Text Summarization

The bar graph demonstrates how well - established techniques like TF - IDF and SVD & Fuzzy Logic perform in Marathi text summarization, attaining higher accuracy levels than more recent approaches like RNN and Deep Learning, which have promise but require improvement. Future developments might concentrate on improving semantic understanding and investigating hybrid strategies to improve performance on tasks involving the summarization of Marathi texts.

5. Conclusion

Text summarizing saves time by compressing input documents and allowing users to easily find important information. Since the 1950s, there has been extensive research on text summarization, but no single technique has consistently produced accurate results. Thus, study is ongoing. So, this paper shows a comparison of different text summarizing approaches for Marathi language. Text summary is primarily done in English, with less emphasis on regional languages such as Marathi. Our findings indicate the need for stronger language models, specifically for Marathi, to increase summarization performance.

Acknowledgement
My sincere gratitude to Dr. Sonal S. Deshmukh, Head of Department, Master of Computer Applications, MGM's JNEC Chhatrapati Sambhajinagar, for their constant support throughout this work. They have given me the inspiration, drive, and support I needed.

References

[1] Akter Sumya, Asa Aysa Siddika, Md Uddin Palash, Hossain Delowar, Roy Shikor Kumar, and Masud Ibn Afjal, “An Extractive Text Summarization Technique for Bengali Document (s) using K - means Clustering Algorithm”, IEEE, 2017.
[2] Ambedkar Kanapala, Pal Sukomal, Pamula Rajendra “Text summarization from legal documents: a survey” © Springer Science+Business Media B. V, 2017.
[3] Awasti Ishitya, Gupta Kunal, Bhoga Prabjot Singhl, Anand Sahejpreet Singh, Soni Piyush Kumar, “ Natural Language Processing (NLP) based Text Summarization - A Survey”, Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT] IEEE, 2021.
[4] Bhoir Shruti, Hule Tanvi, Kadam Deepali, “Marathi Text Summarizer Using Deep Learning Model”, International Research Journal of Engineering and Technology (IRJET) Volume: 08 Issue: 02, 2021.
[5] . Bhosale Shubham, Joshi Diksha, Bhise Vrushali, Deshmukh Rushali, “Marathi e - Newspaper Text Summarization Using Automatic Keyword Extraction Technique”, International Journal of Advance Engineering and Research Development. Volume 5, Issue 03, 2018.
[6] Chaudhari Anishka, Dole Aka, Kadam Deepali, “Marathi text summarization using neural networks”, International Journal of Advance Research and Development, 2019.
[7] Chouk Manasi,, Phadnis Neelam “Text Summarization using Extractive Techniques for Indian Language”, International Journal of Computer Trends and Technology. Volume 69 Issue 6, 44 - 49, 2021.
[8] Dhawale Apurva, Kulkarni Sonali, Kumbhakarna Vaishali “Automatic Pre - Processing of Marathi Text for Summarization”, International Journal of Engineering and Advanced Technology (IJEAT), 2020.
[9] Gaikwad Manisha, Dr. Shinde G. R, Dr. Mahalle Parikshit, Dr. Sable Nilesh and Dr Kharate Namrata, “Automatic Text Summarization: An Extensive Survey”, Grenze International Journal of Engineering and Technology, 2023.

[10] Giri Virat, Dr. Math M. M, Dr. Kulkarni U. P., "Marathi Extractive Text Summarization using Latent Semantic Analysis and Fuzzy Algorithms" Computational Intelligence and Machine Learning Vol - 4, Issue - 1, 2023.
[11] Hajare Utkarsh, Bangade Devendra, Rajgiri Sanket, Dongare Prakash, Khedkar Shilpa, "MARATHI TEXT SUMMARIZATION USING MACHINE LEARNING" IJARIIE - ISSN (O) - 2395 - 4396 - 15995, Vol - 8 Issue - 1, 2022.
[12] Kadam Deepali "A Survey of Extractive Text Summarization for Regional Language Marathi" IRJET, 2021.
[13] Hou Sheng - Luan, Huang Xi - Kun, Fei Chao - Qun, Sun Qi - Lin Wag Chuan - Qing "A Survey of Text Summarization Approaches Based on Deep Learning", Journal of4C Computer Science and Technology 36 (3): 633–663 DOI 10.1007/s11390 - 020 - 0207. Springer, 2020.
[14] Kadam Vaishali, Alazani Samah ali, and Mahender Namarata C "A Text Summarization System for Marathi Languag", Scopus, 2022.
[15] Kakde Kirti, Padalikar H. M., "Marathi Text Summarization using Extractive Technique", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 - 8958 (Online), Volume - 12 Issue - 5, 2023.
[16] Kale Sunil, Mahalle Parikshit, Kachhoria Renu, Kumar Santosh, Chaudhari Prasad, Patil Vivek, "Text summarization through NLP and deep learning mechanism", Journal of Autonomous Intelligence, 2023.
[17] Kamble Satish, Mandage Shivlila, Topale Shubhangi, Vagare Dipali Babbar Prerana, "Survey on Summarization Techniques and Existing Work", International Journal of Applied Engineering Research ISSN 0973 - 4562 Volume 12, Number, 2017.
[18] Kore Rahul, Ray Prachi, Lade Priyanka, Nerurkar Amit "Legal Document Summarization Using Nlp and Ml Techniques", International Journal Of Engineering And Computer Science, 2020.
[19] Kumar Yogesh, Kaur Komalpreet, Kaur Sukhpreet, "Study of Automatic Text summarization approaches in different languages", Springer, 2021.
[20] Mahajan Mayank, Sankhe Sakshi, Shinkar Bhagyashri, Prof. Patil Sainath "Marathi Text Summarizer" Interrnational Journal for Multidisciplinary Research (IJFMR), Volume 6, Issue 3, 2024.
[21] Mridha M. F., Lima Aklima Akter, Nur Kamruddin, Das Sujoy Chandra, Hasan Mahmud, And Kabir Muhammad Mohsin "A Survey of Automatic Text Summarization: Progress, Process and Challenges" DigitalObjectIdentifier10.1109/ACCESS.2021.312978 volume IEEE, 2021/
[22] Rathod Yogeshwari "Extractive Text Summarization of Marathi News Articles", International Research Journal of Engineering and Technology (IRJET) e - ISSN: 2395 - 0056 Volume: 05 Issue: 07, 2018.
[23] Rohil Mukesh kumar, Magotra Varun "An exploratory study of automatic text summarization in biomedical and healthcare domain", Healthcare Analytics, Elsevier 2022.
[24] Sarwadnya Vaishali, Sonawane Sheetal, "Marathi Extractive Text Summarizer using Graph Based Mode" IEEE 2018.
[25] Sharma Bharti, Tomer Minakshi & Kriti Kriti "Extractive text summarization using F - RBM Journal of Statistics and Management System, T& F, 2020.
[26] Shirwandkar Nikhil, Kulkarni Samidha, "Extractive Text Summarization using Deep Learning", IEEE 2018.
[27] Srivastava Akash, Chauhan Kamal, Daharwal Himanshu, Mukati Nikhil, Kavimadan Pranoti Shrikant "Text Summarization using NLP (Natural Langugae Processing) ", IRE Journals, 2022.
[28] Suad Alhojel, Kalita Jugal "Recent Progress on Text Summarization", International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, 2020.
[29] Takale Sheetal Ajayumar, "A Survey of Legal Document Summarization Methods" International Journal of Advanced Research in Computer and Communication Engineering, 2023.
[30] Varagantham Chetana, Reddy Srinija, Yelleni Uday, Kotha Madhumitha, Venkateswara Rao, "Text Summarization Using NLP", Journal of Emerging Technologies and Innovative Research (JETIR), 2022.