e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:06/June-2022 Impact Factor- 6.752 www.irjmets.com

A BRIEF STUDY ON HINDI TEXT SUMMARIZATION USING NATURAL LANGUAGE PROCESSING

Lavanya Pushpakar[*1], Anusha D'souza[*2], Akshaya Bhalikha DS[*3], Sourab Suresh[*4], Ms. Nikhila G[*5]

[*1,2,3,4]Department Of Computer Science, Visvesvaraya Technology University, India

[*5]Asst. Prof. Of AJIET, Department Of Computer Science, Visvesvaraya Technology University, India.

ABSTRACT

The goal of text summarization is to create a summary of the input document that includes important sentences as well as all pertinent information. It removes the unnecessary, insignificant content and provides you with vital information in a compressed format that is usually half the length of the original Input text document. In our approach, we employed the extractive summarization method. The purpose of this strategy is to choose essential sentences from the original input text document and concatenate them into a shorter version. We use numerical approaches to extract the summary in our suggested system, which has the advantage of not requiring any prior data for summary extraction. The most essential sentences that must be included in the summary are determined based on the score of Numerical Features of sentences such as TF-ISF, length of the sentences, the position of the phrases, similarities between the sentences, and Numerical data. Sentences will be chosen based on the score earned by the sentences based on the above-mentioned characteristics. The higher the sentence's score, the more likely it is that the sentence will be included in the summary. The score is determined by extracting features from each sentence in a text document. We have chosen Hindi as the study language for our planned project.

Keywords: Text Summarization, Paragraph, Numerical Method, Hindi Text,Essential Sentences.

I. INTRODUCTION

Every day, an enormous amount of data recirculates over the Internet. For a few years, the Internet has been over-inflated. As a result of all of this, the issue of information or data overload has grown, as has the desire for automatic text summarization. Rather than reading a lengthy paper with numerous theories and examples, the reader will always choose to read a text that is concise yet contains all of the pertinent information. The same goal is served by our proposed study on Hindi text summarization. From the original input text document, the reader receives the most vital and relevant information.

Despite the fact that various ways of summarizing major languages such as English, Swedish, and many other European languages have been proposed. For other languages around the world, text summarization remains a difficult challenge. There are over 1500 languages spoken in India, with 120 major languages spoken throughout the country. The majority of Indians speak Hindi as their first language. Hindi is the official language of Delhi, Chhattisgarh, Himachal Pradesh, Chandigarh, Bihar, Jharkhand, Madhya Pradesh, Haryana, and Rajasthan.

During a related search, it was discovered that various scholars have been working on text summarization in Indian and other languages in recent years. However, there has been very little research into Indian languages. For the pre-processing and processing parts of the document, there are various inherent libraries available in English, but none are available in Hindi. There is a paucity of properly developed Hindi Text Summarization systems, despite the fact that it is widely used in India and adjacent countries. As a result, we've created a technique for condensing Hindi text documents.

Every day, business leaders, specialists, students, and pedagogic advisers must sift through a large number of documents, which takes up a significant amount of time. By extracting the most relevant parts of the papers and making a summary of the entire text, they will be able to save a substantial amount of time, and it will be worthwhile to read only the most important parts. Text summarization can be useful in a variety of situations, such as obtaining news headlines, composing an abstract summary of a lengthy technical document, and writing a book review.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:06/June-2022 Impact Factor- 6.752 www.irjmets.com

II. LITERATURE SURVEY

We conducted a quick study of a range of text summary approaches that have been developed and assessed for the literature review, although largely for English and European languages. A few works in other Indian languages have also been produced.

P. B. Baxendale, "Machine-Made Index for Technical Literature—an Experiment," in IBM Journal of Research and Development, vol. 2, no. 4, pp. 354-361, Oct. 1958.

Much earlier extractive summarizing studies have focused on two key steps: (1) ranking sentences based on a score computed by combining a few or several features such as term frequency (TF), position information, and cue phrases, and (2) picking a few top- ranked sentences to produce a summary. It has been determined that sentences at the start and end of a document are more important than other sentences in the document. Though single nouns and adjectives paired with their frequency of distribution throughout the article do reflect content considerably, the phrase, with its coordinated combination of noun and modifier, shows to be the best index unit of significant terms. [1]

John M. Conroy and Dianne P. O'Leary. 2001. Text summarization via hidden Markov models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). Association for Computing Machinery, New York, NY, USA, 406–407.

They proposed the QR (QR Matrix decomposition) and HMM algorithms for sentence extraction (Hidden Markov Models). The importance of each sentence was judged using the QR method, and the most important sentences were put into the summary. Because some of the remaining sentences were redundant after this, the relative value of the remaining sentences was altered. They went through the procedure again and again until they had caught all of the crucial concepts. The other method was HMM, which is a sequential model for automatic text summarization that determines the chance of each sentence being included in the summary. Only three features were employed in HMM: sentence position, the total number of terms in the sentence, and sentence term similarity in the given document. Finally, the HMM-generated summary was compared to a human-generated summary for evaluation. [2]

H. P. Edmundson. 1969. New Methods in Automatic Extracting. J. ACM 16, 2 (April 1969), 264–285.

To begin, insert cue words into a document that indicate the presence of most indicative words, such as finally, in summary, lastly, and so on. Second, simple feature titles or heading words, for which a sentence was given extra weight if the sentence contained heading words. While earlier research has focused on one aspect of sentence significance, the inclusion of high- frequency content words (keywords), the methods proposed here also address three other aspects: pragmatic words (cue words), title and header words, and structural indicators (sentence location). The research resulted in the development of an operating system as well as a research technique. The steps for compiling the appropriate dictionaries, configuring the control parameters, and comparing theautomatic extracts with manually created extracts are all part of the research approach.[3]

Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph- based lexical centrality as salience in text summarization. J. Artif. Int. Res. 22, 1 (July 2004), 457–479.

Based on a stochastic graph, Erkan and Radev proposed a method for estimating the relative worth of textual units. The LexRank method was used to calculate sentence importance, which is based on the concept of eigenvector centrality in a graph representation of sentences. They look at a novel method for calculating sentence relevance called LexRank, In a graph representation of phrases, this is based on the concept of eigenvector centrality. In most cases, the results reveal that degree-based methods (including LexRank) outperform both centroid-based methods and other DUC systems. In addition, the LexRank with threshold method beats other degree-based algorithms, such as continuous LexRank. They also show that their method is unaffected by data noise, such as that caused byerroneous thematic clustering of papers.[4]

Kupiec, Julian, Jan O. Pedersen, and Francine R. Chen. "A trainable document summarizer." SIGIR '95 (1995).

To summarise is to decrease the complexity and hence the length ofsomething while maintaining some of the original's important traits. The focus of this study is on document extracts, a type of calculated document summary. Even shorter extracts of around20% of the original material can be as instructive as the full text ofa

document, implying that even shorter extracts can be effective suggestive summaries. They talk about how multi-document summarising differs from single-document summarization in that compression, speed, redundancy, and passage selection are all important factors in creating usable summaries. Because the volume and variety of online medical news make it difficult for professionals in the field to read all of it, an automatic multi-document summarising might be useful for a quick review of material on the internet.[5]

H. P. Luhn, "The Automatic Creation of Literature Abstracts," in IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, Apr. 1958.

Automatically generated excerpts of scientific papers and magazine articles that function as traditional abstracts. The machine computes a relative estimate of significance, first for individual words and then for sentences, using statistical information collected from word frequency and distribution. To create summaries, he employed word frequency (the number of times a term appears in a document) and phrase frequency as characteristics. It has long been considered that the most frequently used terms in a document indicate the content's principal theme. Although later research has generated numerous summation methods based on the additional features, (Baxendale, 1958)'s work is still utilized as the foundation for extraction-based summary today.[6]

Kadam, Deepali. "A Comparative Study of Hindi Text Summarization Techniques: Genetic Algorithm and Neural Network." (2015).

Automatic text summarizing is a technique for extracting the most important parts of a source text or texts. It removes the less important, redundant information and replaces it with vital information in a shorter version that is usually half the length of the original text. It may lead you to a solution for the information overload problem in today's environment because it aids in information retrieval. The higher a sentence's score, the more likely it is that it will be included in a summary. They advocate assessing an autonomous text summarizing technique based on sentence extraction using evolutionary algorithms and neural networks to produce a higher quality result. [7]

Wikipedia contributors. (2018). Automatic summarization Wikipedia, the free encyclopedia.

By solving setting complete my email address and devices action complete now by the end then Android smartphone for all active permission setting complete running by all presents the most important or relevant information within the original content, automatic summarization is the process of computationally shortening a set of data to create a subset not allow with action any automatic activities action not active by solving setting complete my email address and devices action complete now by the end then Android smartphone for all active permission setting complete running by all presents the most important or relevant information within the original content.[8]

P. Sethi, S. Sonawane, S. Khanwalker, and R. B. Keskar, "Automatic text summarization of news articles," 2017 International Conference on Big Data, IoT and Data Science (BID), 2017, pp. 23-29.

In academics, text summarization has always been a topic of discussion. They present a text summary technique in this study that focuses on the difficulty of recognizing the most significant parts of a text and generating cohesive summaries. Instead, they produce a summary using a model of text topic progression built from lexical chains. They describe an improved and efficient text summarization system based on lexical chains and the WordNet thesaurus. Furthermore, they implement pronoun resolution and recommend novel scoring strategies to utilize the structure of news stories to overcome the limits of the lexical chain approach to provide a decent summary, overcoming the limitations of the lexical chain approach to generate a good summary.[9]

Rajasekaran, Abirami and Dr R. Varalakshmi. "Review onautomatic text summarization." International Journal of Engineering & Technology (2018): n. pag.

Because there is so much data available in so many different sources and genres, there is a huge need to summarise it for people. In this fast-paced information age, text summarization has grown in prominence. The study of automatically summarising text using various methodologies has exploded in popularity during the last few years. This study examines the many approaches, tactics, and procedures used in Automatic Text Summarization in depth. Automatic Text Summarization (ATS) is a subset of Natural Language Processing (NLP) that involves shortening a source text or a series of text documents/paragraphs while keeping the primary information content. The fundamental goal of Text Summarization is to create a condensed version of

the text that retains all of the important information. [10]

Yadav, A.K., Maurya, A.K., Ranvijay, Ranvijay, R.S. (2021). Extractive text summarization using recent approaches: A survey. Ingénierie des Systèmes d'Information, Vol. 26, No. 1, pp. 109-121.

In this era of rapidly expanding digital media, the volume of text data generated by numerous sources grows by the day and may include complete documents, books, articles, and so on. As a result,we demand strategies and tools that can automatically summarise massive amounts of text data and assist us in determining whether or not they are useful. Text summarising is a method of producinga concise version of a document in the form of a useful summary. Abstractive text summarization and extractive text summarization are two types of text summarization. From the given document, abstractive text summarization provides an abstract type of summary. In extractive text summarising, a summary is constructed from a given source that includes the document's most important sentences. For both types of text summaries, several writers offered numerous strategies. This study gives a survey of graphical-based strategies for extracting text summarization.[11]

P. Janjanam and C. P. Reddy, "Text Summarization: An Essential Study," 2019 International Conference onComputational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6.

In some cases, the proliferation of data from many sources renders individuals incapable of appropriately utilizing information. Text Summarization (TS) is used to get a rapid overview of a large amount of data. The use of linguistics has altered Text Summarization techniques over the decades, resulting in advanced machine learning models. This survey aims to conduct a comprehensive investigation using machine learning, modern graph, and evolutionary-based methods, from feature representationthrough sentence selection and summary generation. The total study will assist researchers in properly handling enormous amounts of data while developing effective Natural Language Processing systems.[12]

R. Boorugu and G. Ramesh, "A Survey on NLP based TextSummarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356. Nowadays, no one can envision living without a smartphone andaccess to the internet. As the use of the internet and cell phones has grown, so has the amount of money spent online. Before orderinganything online, every consumer reads the reviews. However, noteveryone has the ability to read lengthy assessments. As a result,there must be a way to condense large reviews into short sentenceswith few words that convey the same idea. In this case, textsummarization may be useful. Text Summarization is a topic thatmany NLP researchers are interested in. This paper offers anoverview of several text summarising approaches, ranging from themost basic to the most complicated.[13]

P. R. Dedhia, H. P. Pachgade, A. P. Malani, N. Raul, and M. Naik, "Study on Abstractive Text Summarization Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp.1-8.

As a result, an online content summarizer is required to obtain corresponding data as per the application's goal, conveniently and rapidly from various sources of data on the internet. Abstractive Text Summarizer aids in the definition of material by taking into account keywords and in the creation of human-readable summaries. The major goal is to write summaries in such a way that they retain their context. To generate a short summary, several Neural Network models are combined with other machine translation models. The goal of this paper is to highlight and examine current models for abstractive text summarization, as well as to identify topics for future research.[14]

N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5.

It is a proposed method for creating short and precise summariesfor long text documents. Text summarising tackles this challenge by constructing a summary from the document's most significant sentences while preserving the information. In this paper, an extractive text summarising approach for single-document summary is devised and implemented. It selects essential sentences from the text using a combination of Restricted Boltzmann Machine and Fuzzy Logic while keeping the summary meaningful and lossless. Both summaries are then integrated and processed through a series of processes to produce the document's final summary. The results reveal that by generating an appropriate summary, the developed approach addresses

the problem of text overloading.[15]

III. CONCLUSION

Through this brief study, we can design and implement a text summarizer that summarizes a single document in Hindi. For English, a variety of methods and procedures have been devised. However, there are just a few techniques for summarizing Hindi text.When compared to existing algorithms for Hindi text summarization,the performance of our system is satisfactory. Our system's performance may be improved further by refining the stemming process, increasing the amount of statistical and linguistic featuresof a legitimate name, and using the Learning Algorithm to combine features more effectively. In most cases, many reference summaries are examined for each text document during the evaluation process, but we only considered one for each text document. To improve the system in the future, we can include more capabilities like named entity identification, cue words, context information, and word knowledge. Furthermore, this technique can be used for a variety of domains other than news, and the results of these domains can be considered in order to improve the system's overall performance. This technique can be made to work with several documents as well. ANN is a commonly utilized method for sentence ranking (Artificial Neural Network). This method comprises a training phase, during which the system is taught to select the sentences that should be included in the summary.

IV. REFERENCES

[1] P. B. Baxendale, "Machine-Made Index for Technical Literature—An Experiment," in IBM Journal of Researchand Development, vol. 2, no. 4, pp. 354-361, Oct. 1958, DOI: 10.1147/rd.24.0354.

[2] John M. Conroy and Dianne P. O'Leary. 2001. Text summarization via hidden Markov models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01). Association for Computing Machinery, New York, NY, USA, 406–407.

[3] H. P. Edmundson. 1969. New Methods in Automatic Extracting. J. ACM 16, 2 (April 1969), 264–285.

[4] Günes Erkan and Dragomir R. Radev. 2004. LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res. 22, 1 (July 2004), 457– 479.

[5] Kupiec, Julian, Jan O. Pedersen, and Francine R. Chen. "A trainable document summarizer." SIGIR '95 (1995).

[6] H. P. Luhn, "The Automatic Creation of Literature Abstracts," in IBM Journal of Research and Development, vol. 2, no. 2, pp. 159-165, Apr. 1958, DOI: 10.1147/rd.22.0159.

[7] Kadam, Deepali. "A Comparative Study of Hindi Text Summarization Techniques: Genetic Algorithm and Neural Network." (2015).

[8] Wikipedia contributors. (2018). Automatic summarization Wikipedia, the free encyclopaedia from:
https://en.wikipedia.org/w/index.php?title=Automaticsummarizationoldid=822496672([Online; accessed 29- April-2018])

[9] P. Sethi, S. Sonawane, S. Khanwalker, and R. B. Kesar, "Automatic text summarization of news articles," 2017 International Conference on Big Data, IoT and Data Science (BID), 2017, pp. 23-29, DOI: 10.1109/BID.2017.8336568

[10] Rajasekaran, Abirami and Dr R. Varalakshmi. "Review on automatic text summarization." International Journal of Engineering & Technology (2018): n. pag.

[11] Yadav, A.K., Maurya, A.K., Ranvijay, Ranvijay, R.S. (2021). Extractive text summarization using recent approaches: A survey. Ingénierie des Systèmes d'Information, Vol. 26, No. 1, pp. 109-121.
https://doi.org/10.18280/isi.260112

[12] P. Janjanam and C. P. Reddy, "Text Summarization: An Essential Study," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6,
DOI: 10.1109/ICCIDS.2019.8862030.

[13] R. Boorugu and G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product

Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352-356,DOI:10.1109/ICIRCA48905.2020.9183355.

[14] P.R. Dedhia, H. P. Pachgade, A. P. Malani, N. Raul, and M. Naik, "Study on Abstractive Text Summarization Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic- ETITE), 2020, pp. 1-8, DOI: 10.1109/ic- ETITE47903.2020.087.

[15] N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, DOI: 10.1109/ICCUBEA.2018.8697465.