Grenze International Journal of Engineering and Technology, June Issue

A Survey on Text and Voice Summarization of Dravidian Languages

Anjusha Pimpalshende[1], V Baby[2], Challa Sai Venkata Teja[3], Kaluvala Nihal Reddy[4], Kandadi Sai Teja[5] and P Sai Shruthi[6]
[1-6]VNRVJIET, Department of CSE, Hyderabad, India
Email: anjusha_p@vnrvjiet.in, baby_v@vnrvjiet.in, saivenkatateja714@gmail.com, nihalrdycp03@gmail.com, saitejakandadi7@gmail.com, shruthipothula01@gmail.com

***Abstract*— As the majority of India's population is multilingual, there is a growing need to develop efficient techniques for summarizing content in various Indian languages. With a significant portion of the population speaking various Indian languages, there will be an increasing demand for efficient techniques to summarize content and voice in these diverse languages. In recent years, Natural Language Processing (NLP) has emerged as a powerful tool for processing and understanding textual data. The digital revolution has revolutionized education, offering an abundance of educational content in Indian languages. However, this vast amount of information can overwhelm learners, making it challenging for them to grasp and retain essential concepts. To address this challenge, our focus is on leveraging the power of NLP to develop a specialized text and voice summarization system for Indian languages, aiming to optimize the learning experience. The challenges posed by Indian languages are they have diverse scripts, different grammar rules, and regional variations. This linguistic understanding will lay the groundwork for designing NLP algorithms that adapt effectively to different languages, ensuring the accuracy and relevance of the summarized content.**

***Index Terms*— Extractive Summarization, ASR, TextRank, LSA**

I. INTRODUCTION

In the era of information overload, the vast amount of textual and audio data generated every day poses a significant challenge for individuals and organizations seeking to extract valuable insights efficiently. Manual analysis of this deluge of information is time-consuming and often impractical, highlighting the need for automated solutions.
Text summarization, a powerful tool in the digital age, holds significant importance in languages like Telugu, Hindi, and English. First and foremost, text summarization streamlines information retrieval. By condensing lengthy documents, it enables users to quickly pinpoint essential content in search results and databases. Time efficiency is another key advantage, allowing individuals to grasp vital points without investing excessive time.Language accessibility is a critical aspect. Tailoring summarization techniques to specific languages ensures more accurate and contextually relevant summaries. This inclusivity empowers a diverse audience to access crucial information in their preferred languages. Each language has its unique features, making language-specific summarization essential. Summarization simplifies complex topics for learners and students, aiding in knowledge retention. The globalized world benefits from multilingual summarization, breaking language barriers for cross-

Grenze ID: 01.GIJET.10.2.150_1

cultural communication. News agencies utilize it to deliver quick, accurate updates to a wide audience, while legal and regulatory professionals extract key information from lengthy documents for informed decision-making and compliance. Business intelligence relies on summarization to analyze vast amounts of data efficiently, extracting valuable insights. Summarization even offers content personalization, ensuring that readers receive information tailored to their interests and preferences. Summarization with NLP," aims to tackle this challenge by leveraging the power of Natural Language Processing (NLP) to summarize texts and voice recordings in multiple languages. In this survey, "Automated Multilingual Summarization with NLP," we aim to tackle this challenge by leveraging the power of Natural Language Processing (NLP) to summarize texts and voice recordings in multiple languages.This survey paper seeks to provide a comprehensive understanding of the project's importance, the challenges it addresses, and the transformative potential it holds across diverse linguistic, cultural, and professional landscapes.

II. Literature Survey

Sannidhya Raj Chauhan, SateeshkumarAmbesange[1] proposedmethodology which relies focuses on speech summarization of audiobooks without converting them into a transcript using a 1-D convolutional neural network. Acoustic features of the sentence audio is used as input to the model. The output of the model is binary, which tells whether to include that sentence in summary or not .The model achieved an accuracy of nearly 69.57%. It can be further extended to multiple-speaker speech summarization.One of the major advantages of this approach is its language independence. Converting speech summarization into a binary classification task simplifies the problem. One of the limitations is the quality of summaries relies heavily on the accuracy of a specific computer program, and any errors in its understanding of sentences can lead to incomplete or inaccurate summaries.

Aneesh Vartakavi, AmanmeetGarg[2]introduced ASR and extractive text summarization to create audio summaries for podcasts, achieving a 64% accuracy rate. Future work aims to improve accuracy by creating larger datasets for extractive summarization. ASR and extractive text summarization combine to convert audio content into concise text summaries. The system's performance depends greatly on the accuracy of the ASR system, as transcription errors can result in inaccuracies in both text and audio summaries.

Jivin Varghese, PakshalRanawat[3] proposed a project automates text translation, summarization, and speech synthesis to reduce book reading time. The four main modules are Text Extraction (using Vision OCR), Text-to-Speech Synthesis (Tacotron 2 and WaveGlow), Summarization (12-layer abstractive model), and Text-to-Text Translation. The utilization of various machine learning models, including RNN presents a comprehensive approach to text processing. The translation module's capacity to generate semantically accurate Devanagari text is a valuable attribute, particularly beneficial for users seeking precise translations.Limitations involves, the use of a small dataset for English-to-Hindi translation results in difficulties with interpreting uncommon words.

Nisha Vanjari[4] proposed a system involves running Automatic Speech Recognition (ASR) on the podcast to generate a transcript, followed by abstractive text summarization on the transcript. The summary is then converted from text to speech. The system is intended to be accessible via a web application. The use of two benchmark datasets, XSum and SAMSUM, for evaluating the summarization models.Podcast summarization addresses the issue of lengthy and time-consuming podcasts. Issue occurs when podcasts may contain diverse content, making summarization challenging, especially for highly technical or specialized topics.

Dr. Jayashree R[5]used a core methodology involves tagging words with their parts of speech, which aids in grammatical parsing and word sense disambiguation. The research outlines a detailed block diagram of the proposed work, which consists of POS Tagger,ModuleTraining Data Converter,ModuleHMM Trainer Module. POS tagging enhances grammatical parsing, ensuring the generated summaries maintain proper syntax and meaning. The project May not capture the essence of longer texts effectively, as it focuses on individual words or short phrases.

KirusihaBalasundaram[6]presented a approach for text summarization in this research combines both extractive and abstractive techniques. Before Summarizing text document speech that is recorded converted in to text document using google speech recognition API. This text summarization used several modules they are Hierarchical Sentence Representation involves Preprocessing, Tokenization, Word Embedding, Convolutional Sentence Encoder, Bidirectional LSTM-RNN and Sentence Selection using LSTM-RNN.The combination of convolutional sentence encoding and bidirectional LSTM-RNN helps in better understanding the semantic relationships between sentences, leading to more coherent summaries. This model is relatively complex, involving multiple neural network components, which may require significant computational resources for training and inference.

Takatomo Kano[7]introduced an E2E approach, which directly generates summaries from speech signals using a Transformer-based model. It considers alternatives to self-attention mechanisms, such as FNet and Informer, to reduce memory consumption during training. It aligns the embeddings of speech and ASR-generated text using cross-modal attention and generates summaries from this aligned information.Issue related to memory consumption during training, especially for long spoken documents is addressed.

Sujoy Roy, Sumit Mishra[8] designed a model where the tweets are pre-processed and then classified into two categories – informative and non-informative. This model performs Observation from Tweets, Classification Of Microblogging Data, Summarization which involves Summarization Requirement, LSA (LatentSemantic Analysis) Summarization, Luhn Summarization. The model achieved a classification accuracy of 74.268%. In future it can be expanded to filter replies for information and exclude non-relevant tweets to provide better summaries. Support Vector Machine (SVM) with linguistic features is used for classifying tweets into informative and non-informative categories for achieving reasonably high accuracy.Issue occurred because of Microblogging platforms generate a massive volume of data during disasters. Even after filtering, the volume of informative tweets can still be overwhelming.

Aniqa Dilawari[9]employed a approach of two-stage neural network, consisting of an extractor and an abstracter:Extractor: This model selects important sentences from the input article using attention mechanisms.Abstracter: It uses linguistic features such as POS tags, named entity tags (NE), term weights (TF-IDF), and the count of proper nouns to guide the summarization process.The coverage mechanism reduced word repetition, improved the coherence and quality of summaries.The model is limited to smaller datasets. Training the model, especially with large datasets, can be time-consuming and computationally intensive.

Prachi Shah,Nikita P. Desai[10]designed a model which is a Stack based formulation for multiple document summarization. It looks at every solution on that stack. It then tries to extend that solution with every sentence from the document cluster. These extensions are then placed on the stack of the appropriate length. Stacks can efficiently manage a limited amount of content, which can help control memory usage. This model has a limitation i.e stack-based approach produce summaries that miss important information from the input documents, resulting in incomplete coverage.

Prafulla B. Bafna[11]designed a model which performs two tasks Cosine-Based Clustering, Word Cloud Summarization. Cosine-based clustering is a technique used in natural language processing and information retrieval to group similar documents or texts together based on their content. A word cloud is a graphical representation of text data where words are displayed in varying sizes based on their frequency or importance in a given text or set of texts. The use of unsupervised learning and hierarchical clustering efficiently grouped documents with similar content and improved document management. The disadvantage of this approach is it relies on term frequency and does not capture semantic relationships between words.

Sreedhi Deleep Kumar,Sunitha C,Amal Ganesh[12]performed a survey on semantic representation of texts in Indian Languages. For Hindi, WordNet based Rich Semantic graph formulation method. For Telugu, Based on the frequency and ranking method. For Kannada, Combine the GSS and the IDF with TF, Ranking method is used. TF-IDF is widely used in information retrieval systems, such as search engines, to rank documents based on their relevance to a query. TF-IDF representations can result in high-dimensional, sparse vectors, especially in large document collections.

Shubhankita Sood, M. Ferni UkritandRishyParasar[13]analyzes and applies extractive text summarization techniques to Hindi, English, and French text, involving cleaning, tokenization, and sentence scoring to generate concise summaries of the source material. The technique uses various sentence scoring algorithms like frequency-based, Luhn's, cosine similarity, LexRank, TextRank, and Latent Semantic Analysis, offering flexibility for different text types and languages. Using multiple scoring algorithms and tokenization processes can be computationally intensive, potentially impacting scalability, especially with large texts and many sentences.

K Usha Manjari[14] came with TextRank algorithm. PageRank is an algorithm used by Google Search to rank websites in their search engine results similar to page Rank, text rank is also a graph-based ranking algorithm. Instead of webpage links, TextRank algorithms work on words and sentences And Instead of counting incoming links here, the similarity between sentences is used. The F-Measure obtained is 0.621 which is a combined score of precision and recall and is a good quantifier. PageRank algorithm is resistant to spam. Textrank omit the keywords which has lower chance to appear though being meaningful in context.

Geethika JK,Deepamala[15] introduced an extractive approach is used to generate the text summary . This approach has 2 steps- pre processing and processing latent semantic analysis. latent semantic analysis is a powerful analytical tool with the combination of statistical and algebraic methods and this method reveals unseen structure of words ,texts and sentences. LSA Reduces the dimensionality of the text data making it computationally more

efficient while presenting important semantic relationships. LSA struggles to disambiguate words with multiple meanings (polysemy) as it treats each word in isolation.
Dual Translation of International and Indian Regional Language using Recent Machine Translation[16] Identified challenges in translating between international and Indian regional languages. Assessed translation quality. Analyzed scalability and efficiency of machine translation systems. Investigate linguistic and cultural adaptation in Indian regional languages. The focus of the review paper is on utilizing recent machine translation techniques for translating between international languages and Indian regional languages, with an emphasis on the challenges, advancements, and potential solutions in this domain.
Extractive Text Summarisation in Hindi[17]addressed the need for automated text summarization in Hindi, considering the significant growth of data on the web in this language. The focus is on summarizing various types of content, including government data, medical reports, news, and research articles. The paper emphasizes the lack of public datasets for extractive summarization in Hindi and introduces a dataset of 24,253 news articles for evaluation. The primary focus is on extractive text summarization in Hindi, dealing with the challenges posed by the free word order nature of Hindi language. The paper discusses the methodology, including preprocessing, processing, and postprocessing steps, and presents various features used for extractive summarization.
An Automatic Metric for Evaluation Using Embeddings for Machine Translation[18] came with an objective that can develop an automatic metric for evaluating machine translation. Utilized embeddings (distributed representations) for a more nuanced evaluation. Addressed shortcomings of existing evaluation metrics for machine translation. Embeddings are powerful tools in natural language processing, allowing for a more semantic understanding of language. Likely compares the proposed metric with traditional evaluation metrics such as BLEU, METEOR, etc.,
An Efficient English to Hindi Machine Translation System Using Hybrid Mechanism[19] designed a model to overcome the language barrier in India, where different states have various territorial languages. Authors proposed an efficient English to Hindi machine translation system to facilitate communication and understanding among the diverse linguistic communities in India. Introduced a hybrid approach for machine translation, combined declension rules with other techniques like Statistical Machine Translation (SMT) to enhance translation accuracy.
Extractive Text Summarization of Indian Election Commission Manuals[20] proposed a model that can generate shorter versions of the original manuals to provide a brief yet informative document description. Create summaries that are more revealing than just titles, offering insights into the Indian Election Domain manuals. Facilitate rapid assessment of the relevance of the manuals by providing succinct summaries. The paper focuses on implementing an extractive summarization approach for Indian Election Commission Manuals. It involves preprocessing, scoring sentences based on significant words, and selecting top-scoring sentences for concise summaries.

III. METHODOLOGY

LSA

Latent Semantic Analysis (LSA) is not typically used as a preprocessing technique in text summarization, but rather as a core method for extracting semantic relationships and reducing the dimensionality of textual data. It is more commonly employed within the main summarization process. However, LSA can indirectly influence the preprocessing phase by enhancing the quality of the input data.
In a typical text summarization workflow, the initial preprocessing steps include tokenization, stop word removal, stemming, and other techniques aimed at cleaning and structuring the text. After these preparatory steps, you can apply LSA to the preprocessed data to capture semantic information and relationships between words. LSA achieves this by reducing the dimensionality of the data, which helps identify latent semantic patterns.
The benefits of incorporating LSA are most pronounced during the summarization phase, where LSA-generated semantic information can be used to enhance the relevance and coherence of the summary. This can improve the way sentences are ranked or selected, leading to more informative and contextually sound summaries.
In summary, LSA is not a typical preprocessing step but can indirectly influence text summarization by adding a semantic layer to the preprocessing phase. Its primary impact is observed in the core summarization process, where LSA helps create more contextually relevant and coherent summaries based on the semantic information it extracts from the preprocessed text.

Text Rank

TextRank is an algorithm designed to rank the importance of sentences or phrases in a document, making it valuable for constructing effective summaries. However, you can indirectly integrate TextRank into your text

summarization pipeline by using it to calculate sentence similarity. After the initial text preprocessing steps, which may involve tokenization and stop word removal, calculate the similarity between sentences using TextRank or similar techniques like cosine similarity. This information is then used to build a graph where sentences represent nodes, and edges signify their similarity. Applying the TextRank algorithm to this graph ranks the sentences by importance. Finally, the highest-ranked sentences are selected to create the summary, and post-processing steps can be applied to ensure coherence and conciseness. While TextRank is a powerful tool for extracting key content, it's typically used as a post-processing, content-ranking step in text summarization.

Automated Speech Recognition

Automated Speech Recognition (ASR) is not typically used as a preprocessing technique in traditional text summarization, but it can be employed as a valuable initial step when dealing with spoken language data or when you want to summarize spoken content.

In the context of text summarization, ASR can be used to transcribe spoken content into written text. This transcription serves as the basis for subsequent text summarization techniques, as text summarization models typically work with written language. ASR technology converts spoken words into textual form, enabling the summarization algorithm to operate more effectively.

After the ASR step, you can then follow the standard text summarization process, which includes text preprocessing tasks like tokenization, stop word removal, and sentence splitting. Once the spoken content has been transformed into written text, it becomes amenable to the same techniques used for written content summarization.

LexRank

Automated Lexical Rank (LexRank) is typically not used as a preprocessing technique in text summarization. Instead, LexRank is an extractive summarization algorithm that ranks the importance of sentences within a document based on their lexical and structural similarity. It is a core part of the summarization process itself, not a preprocessing step.

However, LexRank can be used in a broader context to assess sentence importance before summarization. In this case, LexRank can be applied to assign importance scores to sentences in a document, and these scores can be used to prioritize sentences during the summarization process. This way, the text summarization model can focus on extracting the most relevant content.

To use LexRank in this way, you would still start with traditional text preprocessing techniques like tokenization and stop word removal, and then apply LexRank to the preprocessed text. The ranked sentences can guide the summarization process, allowing you to select the most significant sentences for inclusion in the summary, which is a valuable enhancement for extractive summarization.

In summary, while LexRank is not a conventional preprocessing technique, it can be integrated into the text summarization pipeline to assess sentence importance and improve the quality of the summary by selecting key sentences.

IV. Research Gap

i. Cross-lingual Voice Summarization: Developing methods for summarizing voice content in one language for an audience that speaks a different language.
ii. Researching ways to provide real-time feedback to summarization algorithms, allowing them to improve and adapt based on user input and preferences.
iii. Addressing privacy concerns related to voice data during summarization, ensuring that sensitive information is not inadvertently shared or stored.
iv. Unlike previous methods restricted to summarizing specific words, our approach excels at summarizing longer paragraphs
v. Investigating techniques for domain-specific summarization, such as medical, legal, or scientific content, which may require specialized knowledge and terminology.

V. Conclusions

In conclusion, this survey paper the study addresses the challenges presented by the diversity of Indian languages, scripts, and regional variations. The survey discusses various methodologies and models proposed for automated summarization, particularly focusing on Indian languages.The paper highlights the significance of text summarization in languages like Telugu, Hindi, and English. Summarization streamlines information retrieval,

saves time, enhances language accessibility, simplifies complex topics, breaks language barriers, and serves various domains, including education, news agencies, legal, and business intelligence. The paper provides insights into the utilization of various techniques, including LSA, TextRank, ASR, LexRank, and other approaches. It outlines how these techniques can be incorporated into the summarization pipeline. The surveyed approaches highlight both the advantages and limitations of different summarization methods. These limitations include language independence, transcript accuracy, computational intensity, and issues with long documents. The survey paper underscores the potential of NLP in addressing the growing need for efficient summarization in Indian languages. It emphasizes the significance of contextually relevant and accurate summarization for a diverse audience. The methodologies discussed in the survey represent various efforts to improve text and voice summarization, while acknowledging the challenges posed by Indian languages.

REFERENCES

[1] Chauhan, Sannidhya Raj, SateeshkumarAmbesange, and Shashidhar G. Koolagudi. "Speech Summarization Using Prosodic Features and 1-D Convolutional Neural Network." 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE). Vol. 7. IEEE, 2022.

[2] Vartakavi, Aneesh, Amanmeet Garg, and Zafar Rafii. "Audio summarization for podcasts." 2021 29th European signal processing conference (EUSIPCO). IEEE, 2021.

[3] Varghese, Jivin, et al. "Audio Synthesis Translation and Auto-Summarization (ASTA)." 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT). IEEE, 2022.

[4] Vanjari, Nisha, et al. "Podcast Summarization System." 2022 5th International Conference on Advances in Science and Technology (ICAST). IEEE, 2022.

[5] Jayashree, R., Basavaraj S. Anami, and B. K. Poornima. "Text Document Summarization Using POS tagging for Kannada Text Documents." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021.

[6] Balasundaram, Kirusiha, and C. R. J. Amalraj. "Speech document summarization using neural network." 2019 4th International Conference on Information Technology Research (ICITR). IEEE, 2019.

[7] Kano, Takatomo, et al. "Speech Summarization of Long Spoken Document: Improving Memory Efficiency of Speech/Text Encoders." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[8] Roy, Sujoy, Sumit Mishra, and Rakesh Matam. "Classification and summarization for informative tweets." 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2020.

[9] Dilawari, Aniqa, et al. "Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space." IEEE Access 11 (2023): 23557-23564.

[10] Shah, Prachi, and Nikita P. Desai. "A survey of automatic text summarization techniques for Indian and foreign languages." 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE, 2016.

[11] Bafna, Prafulla B., and Jatinderkumar R. Saini. "Hindi multi-document word cloud based summarization through unsupervised learning." 2019 9th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-19). IEEE, 2019.

[12] Kumar, SreedhiDeleep, C. Sunitha, and Amal Ganesh. "Semantic representation of texts in Indian languages—A review." 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018.

[13] Sood, Shubhankita, M. Ferni Ukrit, and Rishy Parasar. "Analysis of Extractive Text Summarisation Techniques for Multilingual Texts." 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2023.

[14] Manjari, K. Usha. "Extractive summarization of Telugu documents using TextRank algorithm." 2020 Fourth international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC). IEEE, 2020.

[15] Geetha, J. Kamala, and N. Deepamala. "Kannada text summarization using latent semantic analysis." 2015 International conference on advances in computing, communications and informatics (ICACCI). IEEE, 2015.

[16] Jayanthi, N., et al. "Dual translation of international and Indian regional language using recent machine translation." 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2020.

[17] Vijay, Sakshee, et al. "Extractive text summarisation in hindi." 2017 International Conference on Asian Language Processing (IALP). IEEE, 2017.

[18] Mukherjee, Ananya, et al. "Mee: An automatic metric for evaluation using embeddings for machine translation." 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2020.

[19] Nair, Jayashree, K. Amrutha Krishnan, and R. Deetha. "An efficient English to Hindi machine translation system using hybrid mechanism." 2016 international conference on advances in computing, communications and informatics (ICACCI). IEEE, 2016.

[20] Shukla, Seema, et al. "Extractive Text Summarization of Indian Election Commission Manuals." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.

[21] Pimpalshende, Anjusha, and A. R. Mahajan. "Test model for stop word removal of devnagari text documents based on finite automata." 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). IEEE, 2017.
[22] Pimpalshende, A. N. "Overview of text summarization extractive techniques." International Journal of Engineering and Computer Science 2.4 (2013): 1205-1214.
[23] Pimpalshende, A., and A. R. Mahajan. "Extraction of root words using morphological analyzer for Hindi text." Int J Soft Comput 13.5 (2018): 134-138.
[24] Pimpalshende, Anjusha, and A. R. Mahajan. "Pre-processing phase of Hindi language text summarization System." International Journal of Computer Science and Information Security (IJCSIS) 14.5 (2016).
[25] Potnurwar, A., Pimpalshende, etl (2020). Extractive multi-document text summarization by using binary particle swarm optimization. Helix, 10(04), 263-265.