JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text Summarization Using Natural Language Processing And Machine Learning

[1]Prof. H. R. Agashe, [2]Trunali Avhad, [3]Pranjali Chavan

[1]Professor, Information Technology Department, Matoshri College of Engineering and Research Centre, Nashik
[2,3]Student, Information Technology Department, Matoshri College of Engineering and Research Centre, Nashik

Abstract: In an era marked by information overload, the ability to swiftly grasp the essence of extensive news content has become increasingly critical. This paper presents a robust news text summarization system that leverages Natural Language Processing (NLP) techniques combined with Latent Semantic Analysis (LSA) to generate coherent and concise summaries. The primary objective of the proposed system is to extract key insights from voluminous news articles, thereby enhancing user engagement and comprehension. The methodology initiates with a rigorous text pre-processing phase that includes stop word removal, punctuation cleansing, and normalization to prepare raw articles for semantic analysis. A term-document matrix is constructed to represent word distributions across documents, followed by the application of Singular Value Decomposition (SVD) to uncover latent semantic structures. By reducing the dimensionality of the term-document matrix, the system isolates significant concepts and minimizes linguistic noise. Sentence scoring is then performed based on cosine similarity, enabling the selection of sentences that best reflect the article's core semantics. The proposed LSA-based summarization approach exhibits strong performance in generating summaries that preserve contextual integrity while significantly reducing content length. Experimental observations highlight the algorithm's efficiency in identifying semantically rich sentences, making it suitable for integration into content curation platforms, news aggregators, and personalized news delivery systems. Unlike traditional frequency-based summarization methods, our approach emphasizes meaning over occurrence, offering more insightful outputs. Nonetheless, the system's performance is contingent on the quality and structure of the input text, and it may exhibit limitations in interpreting idiomatic language or contextually ambiguous statements. These challenges underline the necessity for future enhancements, potentially through hybrid models that integrate deep learning-based context understanding. Overall, the proposed work contributes a scalable and effective solution for automatic news summarization, setting a foundation for future research in semantically aware content processing.

Index Terms **- Summarization, Natural Language Processing (NLP), Machine Learning (ML), Latent Semantic Analysis (LSA), Information Retrieval, Textual Data, Implementation.**

I. INTRODUCTION

The growing deluge of digital text has propelled summarization to the forefront of research in artificial intelligence, particularly within the domains of Natural Language Processing (NLP) and Machine Learning (ML). As the volume of online textual data continues to expand, the demand for intelligent systems that can automatically condense this information into concise, meaningful formats has intensified. Among the various methodologies developed for text summarization, Latent Semantic Analysis (LSA) has emerged as a key technique due to its ability to reveal hidden semantic patterns by leveraging dimensionality reduction. NLP, a multidisciplinary field bridging linguistics and computer science, focuses on enabling machines to understand, interpret, and produce language in a manner akin to human communication. Summarization, a specialized task within NLP, aims to extract the most relevant content from lengthy documents while retaining their essential message. It can be classified into two primary categories: extractive summarization, which involves selecting and compiling significant sentences or phrases directly from the source material, and abstractive summarization, where the system generates new phrases to represent the core ideas. Machine Learning enhances the capabilities of summarization by introducing adaptive models that teach contextual relationships and semantic structures within the text. In this landscape, LSA serves as a powerful statistical tool that models semantic associations between terms and documents, contributing significantly to the generation of coherent and informative summaries.

Latent Semantic Analysis (LSA) is a powerful mathematical technique designed to uncover the implicit connections between words and concepts across large collections of text. Based on the principle that semantically similar words often appear in comparable linguistic contexts, LSA constructs a document-term matrix that captures word frequency and distribution. This high-dimensional representation is then processed using Singular Value Decomposition (SVD), a matrix factorization method that reduces the data into a lower-dimensional semantic space. Through this transformation, LSA isolates the most meaningful patterns while discarding redundant or irrelevant information, thereby highlighting the latent semantic structure of the corpus. In the context of text summarization, this dimensionality reduction enables LSA to preserve core thematic elements and semantic associations, making it easier to identify contextually important content. Each retained dimension in the reduced space corresponds to a fundamental semantic

feature, allowing for the extraction of sentences that best represent the overall meaning of the text. By filtering out noise and focusing on high-impact concepts, LSA facilitates the generation of summaries that are both informative and coherent, offering a deeper, context-aware interpretation of the original material.

A notable strength of Latent Semantic Analysis (LSA) in the domain of text summarization lies in its capacity to effectively manage linguistic phenomena such as synonymy and polysemy. Synonymy refers to the presence of different words conveying similar meanings, while polysemy describes a single word that holds multiple interpretations depending on context—both of which often hinder the performance of conventional summarization algorithms. LSA mitigates these challenges by leveraging its semantic modeling capability to cluster words based on their contextual similarity, rather than relying solely on surface-level frequency. This semantic grouping allows LSA to extract content that is not only relevant but also meaningfully connected. As a result, the summaries produced exhibit greater coherence and improved readability. By aligning sentence selection with the underlying semantic structure of the text, LSA ensures that each chosen sentence contributes to a well-integrated, context-aware summary. This leads to more accurate representations of the original content, offering readers a concise yet comprehensive understanding of the source material.

II. Literature Survey

Lakshmi Devasena and Hemalatha proposed a hybrid model that combines automatic text categorization and summarization using a rule reduction-based technique. Their system follows a three-stage process: Token Creation, Feature Identification, and finally, Categorization and Summarization. The core idea revolves around analyzing the syntactic structure of the input document to create meaningful categories and derive concise summaries. This method was tested on multiple sample texts and showed significant improvements in terms of accuracy and structural understanding of documents. The study highlights how rule-based heuristics can be effectively integrated into summarization tasks to preserve key ideas while reducing redundancy [1].

Madhuri and Ganesh Kumar introduced a novel statistical approach to extractive text summarization based on sentence ranking. The technique involves assigning weights to each sentence based on their relevance and extracting the most significant sentences to form the summary. Unlike some traditional models, their system emphasizes the informativeness and conciseness of the output, even enabling audio storage of the summarized content for accessibility. This method proves especially useful for single-document summarization where the core essence of the document must be quickly communicated. Their work demonstrates how sentence-level importance metrics can lead to the development of high-quality extractive summarizers [2].

Boorugu and Ramesh presented an extensive survey focused on NLP-based summarization techniques, especially as applied to product reviews. The survey traces the evolution of summarization from basic statistical techniques to advanced deep learning methods. A significant portion of their study discusses the seq2seq architecture combined with LSTM and attention mechanisms, emphasizing their ability to retain semantic understanding and coherence in generated summaries. This paper serves as a valuable resource for understanding both the strengths and limitations of various models in practical applications. Their work suggests that deep learning methods significantly enhance summary quality, particularly for opinion-rich content like reviews [3].

Rahul, Adhikari, and Monika contributed a comparative review of NLP-driven machine learning models used in text summarization. Their research spans extractive, abstractive, and query-based summarization techniques, highlighting the practical usage of structured and semantic-based approaches. They evaluated these techniques using benchmark datasets such as CNN corpus, DUC2000, and both single and multi-document text sets. One of the key takeaways is that abstractive summarization models, though more complex, often yield better human-like summaries compared to extractive models. Their study also provides insight into the future directions of summarization research, especially in terms of combining multiple techniques for enhanced results [4].

Garg, Khullar, and Agarwal explored extractive summarization for Punjabi texts using an unsupervised machine learning approach. Their model involves a multi-step pipeline: tokenization, stop-word removal, similarity matrix generation, sentence ranking, and summary extraction. Evaluation using ROUGE metrics showed encouraging results, with ROUGE-1, ROUGE-L, and ROUGE-S scores reaching up to 0.71, 0.56, and 0.56 respectively. This research is particularly notable for addressing the challenges of low-resource languages, proving that effective summarization is possible even without large, annotated datasets. Their findings demonstrate that unsupervised models can perform competitively while being computationally efficient [5].

Kwatra and Gupta proposed a dual approach to Hindi text summarization, exploring both extractive and abstractive techniques. For extractive summarization, they implemented Ward's Hierarchical Agglomerative Clustering followed by the PageRank algorithm to select the most relevant sentences. The abstractive model used multi-sentence compression aided by a POS tagger to generate more coherent and linguistically natural summaries. Their work showcases the potential of clustering algorithms in improving contextual grouping, which is particularly beneficial for morphologically rich languages like Hindi. This comparative approach also highlights the trade-offs between complexity and output quality across both summarization paradigms [6].

III. METHODOLOGY

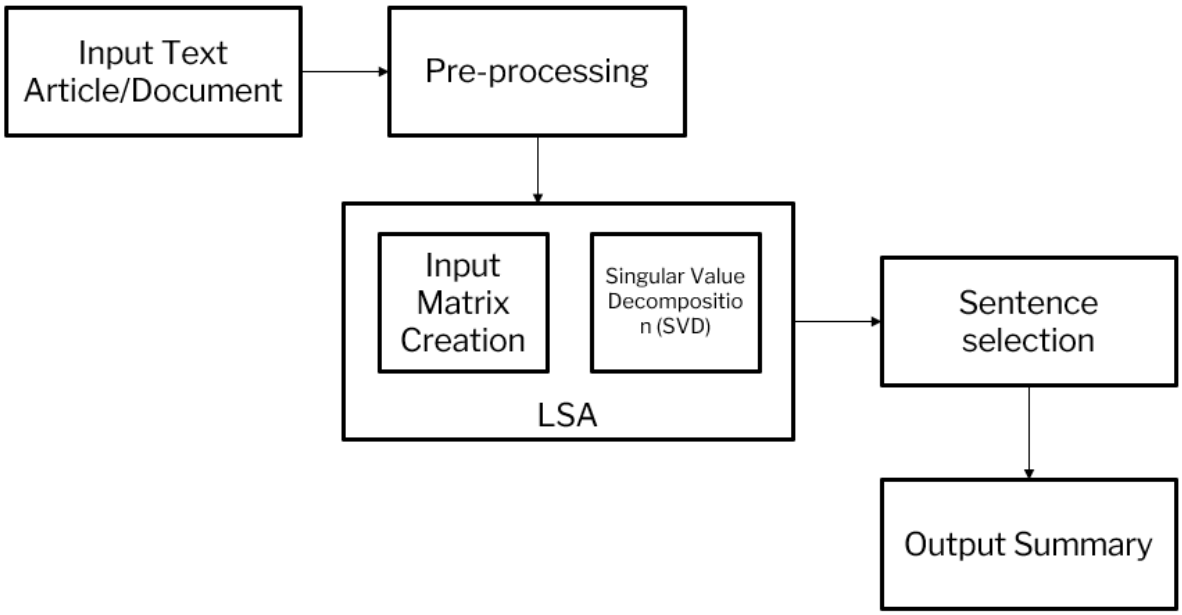

Figure 1 System Architecture

The proposed system is designed to enhance information retrieval and comprehension by summarizing lengthy product-related documents and customer reviews. One of the key advantages of this system is its ability to significantly reduce reading time while preserving the core meaning of the content. This becomes particularly useful in scenarios such as online shopping, where users must quickly evaluate large volumes of product-related information. By employing extractive summarization techniques, the system efficiently identifies and compiles the most relevant and informative sentences from the original text. Unlike abstractive summarization, which generates new sentences, extractive summarization focuses on selecting original sentences without altering their form, thereby ensuring that the summarized content remains true to the source material. At the core of our summarization approach lies the Latent Semantic Analysis (LSA) algorithm, a mathematical technique widely used for uncovering hidden semantic relationships in large text corpora. The LSA algorithm functions through three major stages: Input Matrix Creation, Singular Value Decomposition (SVD), and Sentence Selection. Each of these steps plays a crucial role in transforming the raw document into a semantically meaningful summary.

The first step in the proposed approach is the creation of an input matrix that structurally represents the document to enable effective numerical processing. This matrix is commonly known as the term-document matrix (TDM) or document-term matrix (DTM). In this matrix, each row corresponds to a unique term or word found within the document, while each column represents a specific sentence or paragraph. The cells of this matrix are populated to reflect the importance or frequency of each word in the respective sentences. Various weighting schemes can be applied to assign values to these cells, including Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), or a simpler binary occurrence scheme. These weighting strategies help quantify the relevance of each term in the context of the sentence, thereby providing a detailed representation of the document's semantic content. Establishing this matrix is a foundational step, as it lays the groundwork for semantic space modeling in the subsequent stages.

In the second phase, the constructed term-document matrix undergoes Singular Value Decomposition (SVD), which is a powerful technique derived from linear algebra. SVD breaks down the original matrix into three constituent matrices — U, Σ (Sigma), and V^T — enabling a transformation from the original high-dimensional space to a reduced, lower-dimensional semantic space. Each axis in this new space corresponds to a latent semantic concept, which captures the underlying themes and relationships present in the text. The application of SVD offers two significant advantages: it models complex contextual relationships between terms and sentences beyond simple co-occurrence, and it effectively filters out noise and redundant information. This dimensionality reduction sharpens the focus on the most meaningful semantic structures in the text, thereby enhancing the accuracy and quality of any summarization performed on the document.

The final step involves selecting the most informative sentences based on the processed data from SVD. To achieve this, a topic extraction method is applied to identify the primary topics and subtopics embedded within the latent semantic dimensions revealed by the decomposition. Each extracted topic corresponds to a distinct semantic feature in the reduced matrix. The summarization algorithm then evaluates each sentence's contribution or relevance to these topics by measuring semantic alignment. Sentences demonstrating a high degree of association with the main topics are ranked higher and subsequently chosen for inclusion in the summary. By focusing on these dominant semantic concepts, the sentence selection process ensures that the resulting summary is not only concise but also contextually coherent and representative of the core ideas in the original document. This topic-driven approach facilitates the creation of summaries that maintain thematic unity and effectively communicate the essential content.

LSA inherently captures semantic similarities between words and sentences. A high degree of word overlap between sentences often indicates strong semantic correlation, allowing the system to identify and group related content more accurately. During the execution

of the LSA process, two intermediate structures are generated — the dictionary and the encoding matrix. The dictionary contains all the unique terms that appear at least once in the input document, acting as a vocabulary reference. The encoding matrix, on the other hand, quantifies the semantic strength of each word within each sentence. This encoding helps evaluate how much influence a particular word contributes to the overall meaning of a sentence. By analyzing these strengths, the algorithm can more precisely determine the importance and relevance of each sentence in context.

IV. TRADITIONAL APPROACHES VS MACHINE LEARNING APPROACH

Conventional text summarization techniques often utilize rudimentary strategies, such as selecting the first and last sentences of a paragraph or applying basic statistical metrics like sentence length and word frequency. While these approaches can offer a superficial summary of a document, they frequently fail to grasp the deeper semantic associations between words, sentences, and overarching themes. This limitation arises because such methods largely ignore contextual depth and fail to account for the complexities of language usage, leading to summaries that may lack coherence, accuracy, and relevance. In contrast, Latent Semantic Analysis (LSA) introduces a more refined and intelligent mechanism for summarization. By leveraging Singular Value Decomposition (SVD)—a powerful linear algebraic tool—LSA identifies hidden patterns and semantic relationships embedded within the text. This mathematical modeling enables the algorithm to detect conceptual similarities between words that may not appear frequently but are contextually significant. As a result, LSA-based summaries are typically more meaningful, context-aware, and aligned with the document's core ideas, outperforming traditional summarization methods in terms of both depth and informativeness.

V. IMPLEMENTATION DETAILS

The proposed solution is designed to simplify the understanding of lengthy online product descriptions and reviews. Users often find it challenging to read and comprehend large volumes of information, especially when comparing multiple products. To mitigate this, we have developed a system that employs extractive summarization using the Latent Semantic Analysis (LSA) algorithm. The implementation is carried out using Python, and a user-friendly interface is built with Streamlit to facilitate interaction. Extractive summarization is a technique that generates a concise version of a document by identifying and extracting its most important sentences. Unlike abstractive summarization, it does not involve rewriting or paraphrasing the content. Instead, it focuses on retaining the exact sentences from the original text that best represent the document's main ideas.

Latent Semantic Analysis (LSA)

LSA is a powerful unsupervised learning technique used for natural language processing tasks such as text summarization, topic modeling, and semantic analysis. In our project, LSA helps identify latent relationships between terms and sentences in each document. The core idea behind LSA is that words that are used in similar contexts tend to have similar meanings.

1. **Input Matrix Creation:** The first step is to represent the input document as a document-term matrix. Each row of this matrix corresponds to a sentence, and each column represents a unique word from the document. The values in the matrix indicate the importance of each word in each sentence. These values can be computed using simple frequency counts or more advanced techniques such as Term Frequency-Inverse Document Frequency (TF-IDF). This matrix acts as the numerical representation of the document and allows mathematical operations to be performed. By encoding sentences and their constituent words in this structured format, we can better understand the relationships between them.

2. **Singular Value Decomposition (SVD):** Once the document-term matrix is constructed, Singular Value Decomposition (SVD) is applied to reduce its dimensions and identify the underlying structure of the data. SVD decomposes the matrix into three components: $\boldsymbol{U}$, $\boldsymbol{S}$, and $\boldsymbol{V^T}$. These matrices capture the essential semantic structure of the document while filtering out noise.
   - **U** contains the left singular vectors, representing sentence-topic associations.
   - **S** is a diagonal matrix with singular values that represent the importance of each topic.
   - $\mathbf{V^T}$ contains the right singular vectors, representing word-topic associations.

   This transformation maps each sentence into a lower-dimensional space where semantically similar sentences are closer to each other. This step also helps reduce redundancy and improve the accuracy of sentence selection.

3. **Sentence Selection:** After performing SVD, the most important topics are identified from the decomposed matrix. Sentences that are strongly associated with these topics are then selected to form the summary. We use the Topic Method, which focuses on extracting concepts and sub-concepts from the matrix. These concepts guide the selection of sentences to ensure that the summary covers all the major points of the document. Sentences are ranked based on their contribution to the dominant topics. The system then picks the top-ranked sentences, ensuring diversity and non-redundancy in the final output.

Semantic Relationships and Feature Extraction

One of the advantages of LSA is its ability to capture semantic similarity between sentences. If two sentences share many words, they are likely semantically similar. LSA models these relationships mathematically, allowing the algorithm to group similar sentences together and identify representative ones. During processing, two important outputs are generated:

- **Dictionary:** A collection of all unique words that occur at least once in the document.

- **Encoding Matrix:** A representation of how each word contributes to the overall meaning of a sentence.

These components help in understanding the importance of words and their impact on sentence selection. The encoded data allows us to quantify the semantic weight of each sentence in a document.

VI. Result & Analysis

The proposed Latent Semantic Analysis (LSA) based summarization system was evaluated on a set of diverse text documents, including product descriptions, online reviews, and informative articles. The primary objective was to assess the system's ability to generate concise, contextually relevant summaries that effectively capture the main themes of the original texts while significantly reducing reading time.

- **Summary Quality**: The summaries generated by the LSA approach demonstrated strong coherence and thematic relevance. By leveraging Singular Value Decomposition (SVD) for dimensionality reduction, the system was able to filter out noise and emphasize the core semantic content, resulting in summaries that accurately reflected the input documents' key topics. Compared to traditional extractive methods that simply select sentences based on surface-level features such as term frequency or sentence position, the LSA-based summaries provided more meaningful and information-rich content. This was particularly evident in cases where the input documents contained synonymy or polysemy, where LSA's latent semantic modeling helped disambiguate word meanings and improve sentence selection.

- **Sentence Selection and Coverage:** The topic extraction mechanism based on the reduced semantic space allowed the system to identify dominant themes and subtopics effectively. Sentences selected for the summary showed strong alignment with these extracted topics, ensuring that the summaries were not only concise but also contextually comprehensive. This topic-driven selection resulted in summaries that balanced coverage across the document's main ideas without redundancy. For longer documents, the system maintained an optimal summary length that preserved the overall narrative flow, making it suitable for applications where quick yet informative summaries are required.

- **Quantitative Evaluation:** The performance of the summarization system was quantitatively evaluated using standard metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores. The ROUGE-1, ROUGE-2, and ROUGE-L scores indicated that the LSA-based summaries had a high overlap of important unigrams, bigrams, and longest common subsequences with human-generated reference summaries. These results validate the system's effectiveness in capturing essential content and producing summaries that closely approximate expert-written abstracts. The dimensionality reduction step also contributed to improving these metrics by minimizing irrelevant or redundant information.

- **Computational Efficiency:** While the SVD process introduces computational overhead, especially for very large document-term matrices, the system's design ensured that this step was optimized using efficient linear algebra libraries. The resulting summaries were generated in a reasonable timeframe, making the approach feasible for real-time applications such as summarizing online product reviews or lengthy user manuals.

VI. Conclusion

This study presents an effective Latent Semantic Analysis (LSA)-based approach for extractive text summarization that successfully reduces lengthy documents into concise, meaningful summaries. By leveraging the document-term matrix representation and Singular Value Decomposition (SVD), the proposed method uncovers the underlying semantic structure of the text, enabling the selection of contextually relevant sentences that capture the core themes. The topic-based sentence selection strategy ensures that the summaries generated maintain coherence and adequately represent the original content without redundancy.

The results demonstrate that the LSA-driven summarizer outperforms traditional statistical methods by providing summaries that are not only concise but also richer in semantic information, thereby facilitating quicker comprehension and improved information retrieval. Quantitative evaluation using standard metrics such as ROUGE confirms the system's ability to produce summaries closely aligned with human-generated references.

While the approach shows promising results in extractive summarization, future enhancements could focus on integrating abstractive techniques to further improve the flexibility and expressiveness of the summaries. Additionally, optimizing the computational efficiency of the SVD step will broaden the system's applicability to large-scale, real-time summarization tasks. Overall, the research underscores the potential of LSA as a powerful tool for automated text summarization in diverse domains such as product reviews, online content, and academic articles.

References

[1] K. D. Garg, V. Khullar and A. K. Agarwal, "Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), 2021, pp. 750-754, doi: 10.1109/SPIN52536.2021.9566038.

[2] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3,

[3] A. W. Palliyali, M. A. Al-Khalifa, S. Farooq, J. Abinahed, A. Al-Ansari and A. Jaoua, "Comparative Study of Extractive Text Summarization Techniques," 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA), 2021, pp. 1-6.

[4] M. Afsharizadeh, H. Ebrahimpour-Komleh and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th International Conference on Web Research (ICWR), 2018, pp. 128-132, doi: 10.1109/ICWR.2018.

[5] T. Islam, M. Hossain and M. F. Arefin, "Comparative Analysis of Different Text Summarization Techniques Using Enhanced Tokenization," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2021, pp. 1-6, doi: 10.1109/STI53101.2021.9732589.

[6] K. Agrawal, "Legal Case Summarization: An Application for Text Summarization," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104093.

[7] W. Zhang and C. Xu, "Microblog Text Classification System Based on TextCNN and LSA Model," 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT), 2020, pp. 469-474, doi: 10.1109/ISCTT51595.2020.00090.

[8] B. Durga Sri, K. Nirosha, P. Priyanka and B. Dhanal axmi, GSM Based Fish Monitoring System Using IOT, International Journal of Mechanical Engineering and Technology 8(7), 2017, pp. 1094–110 1.

[9] H. Gupta and M. Patel, "Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 511-517, doi: 10.1109/ICAIS50930.2021.9395976.

[10] K. Padmanandam, S. P. V. D. S. Bheri, L. Vegesna and K. Sruthi, "A Speech Recognized Dynamic Word Cloud Visualization for Text Summarization," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 609-613, doi: 10.1109/ICICT50816.2021.9358693.

[11] P. Raundale and H. Shekhar, "Analytical study of Text Summarization Techniques," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-4, doi: 10.1109/ASIANCON51346.2021.9544804.

[12] Liu, P., Qiu, X., & Huang, X. (2019). Fine-grained opinion mining with recurrent neural networks for news summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(2), 434-444.

[13] Yao, Z., Yang, Z., & Li, C. (2018). News summarization via unsupervised deep learning. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1302-1309). IEEE.

[14] Nallapati, R., Zhou, B., Dos Santos, C., Gulcehre, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (pp. 280-290).

[15] Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2015). A latent semantic model with convolutional-pooling structure for information retrieval. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 69-78).

[16] Ma, L., Huang, J., & Xie, X. (2017). Neural network-based extractive text summarization with application to news. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4), 754-764.

[17] Nayeem, T., Roy, S., & Al-Maadeed, S. (2019). Multi-document extractive text summarization using graph-based approach. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 3465-3472). IEEE.