# A Survey on TextSummarization of News Articles Using Natural Language Processing

Prof. Ghadge S.V.[1,] Prof Shah S.N.2, Kiran Narute[3], Aditya Ranaware[4], Saurav Nagawade[5]
[1]*Professor, Department of Computer Engineering, Sharadchandra Pawar college of Engineering and Technology Someshwarnagar*
[2]*HOD, Department of Computer Engineering, Sharadchandra Pawar college of Engineering and Technology Someshwarnagar*
[3,4,5]*Student, Department of Computer Engineering, Sharadchandra Pawar college of Engineering and Technology Someshwarnagar*

***Abstract*—This approach utilizing PageRank algorithm for summarization is in fact extractive in nature and takes inspiration from the original rank ordering done by Google. Given the exponential increase in load of news, summarization has become the need of the hour to make a quick understanding about significant information. The process involves preprocessing along with embedding the sentences through Word2Vec, followed by capturing real important points and ranking them up using PageRank. The results show that this technique well extracts and ranks important sentences concisely without compromising on the theme; hence, suitable for real-time media applications.**

***Index Terms*—Text Summarization,PageRank, Extractive Summarization, News Summarization, Word2Vec, Sentence Ranking, NLP.**

## I. INTRODUCTION

In today's digital world, people are constantly inundated with an endless stream of text from various sources, be it news websites, social media, or research articles. However, it becomes nearly impossible to extract key points from this barrage of information. Thus arises the need for efcient means to summarize text so that one can use it for extracting important insights without having to read them in detail. Natural Language Processing, or NLP in short, has a major contribution in this regard, and text summarization is by far one of the most important applications of NLP. It condenses large volumes of information into bullets, hence making them easy to read and decide upon. This paper takes into consideration the PageRank algorithm, which is a graph-based ranking approach, in extractive text summarization. First, unlike the earlier methods which merely consider word frequency to defne important sentences, PageRank constructs a document as a network of sentences. Each sentence will then be considered as a node, and edges between sentences are created via the establishment of similarity or relation between those sentences. The logic of the algorithm is that more the interconnectedness of sentences, the more it would attach importance to sentences and as such select them into the summary of the relevant ones. A summary is created such that it keeps the original context and logic intact while keeping the important information by examining the structure and connections among the sentences. PageRank thus provides a more effective method to access relevant content than ordinary keyword matching. We will probe into different aspects of how PageRank can improve text summarizedby users' efciency on processing huge volumes of textual data. It is quite possible to improve PageRank for summarization by integrating various other NLP techniques, such as Word2Vec and the use of sentence embeddings, because those techniques involve comprehending thesemantic interrelations between sentences, rather than simply their structural connection. Acombination of Word2Vec and PageRank would provide for sentences to be ranked according not only to their placement within a given text or document, but to meanings and relevance.This will not only increase the informativeness and diversity of the summaries, hence ensuring inclusion of key ideas from the various sections of the text, but will also guarantee the addition of text-to-speech (TTS)

conversion into the summarization so users can listen to the resulting summaries instead of reading them. This comes in handy for systems that further include news briefngs, educational tools, and accessibility contexts. With its combination of graph-based ranking and deep learning techniques, text summarization will make for effective and user-friendly solutions.

## II. LITERATURE SURVEY

The automation of text summaries has improved largely due to Natural Language Processing and the modeling of machine learning techniques. Therefore, several methods for extractive and abstractive summary generation have been explored with the aim of improving efficiency, coherence, and relevance in generated summaries. An analysis has been done by [Yadav et al. (2021)] on extractive summarization techniques comparing statistical models, deep learning approaches as well as hybrid models. From their study, it was attested that machine learning-based ranking methods work better than traditional approaches in filtering relevant content. Out of this, our study has dissolved the concept of hybrid ranking model within PageRank based approach so as to facilitate better sentence selection through combining statistical ranking with NLP embeddings. Sentences ranking models and hybrid extractive summarization were proposed by [Qaroush et al. (2021)] to improve sentence ranking through both semantic and statistical features. The model emphasizes context-aware selection so that the extracted summaries remain coherent. We adapted the model by implementing contextual sentence scoring into our PageRank algorithm, thereby refining sentence importance judgment within the text. [Wazery et al. (2022)] used a deep-learning-based abstractive summarization approach employing transformers and attention mechanisms to ensure improvement in summary quality. The self-attention mechanism is clearly shown to maintain coherence. While predominantly relying on extractive summarization, our system drew inspiration from this work to examine the possibility of incorporating transformer-based techniques in the future for improved summary fuency. Using k-medoid clustering and evolutionary optimization techniques, they developed a multidocument summarization model that favors summary diversity and coverage. The essence of grouping similar sentences by clustering before applying PageRank was included in our study, which helped avoid redundancy and maintain balanced representation of the source text.[Suleiman et al. (2020)] presented some challenges in the area of abstractive summarization concerning dataset selection, evaluation metrics, and text generation techniques, and thus give guidelines on judgment on the summary quality with regards to the right dataset and evaluation methods to be employed. From their works, we adopted a more structured evaluation for the effectiveness of our PageRank-based summarization model and thus could ensure better evaluation based on ROUGE scores. These research works are advancements that emphasize extractive and abstractive summarizations with respect to improvements through deep learning and clustering, which hybrid approaches also initiate. Our system proposes improvements in methods by combining PageRankbased sentence ranking with contextual NLP techniques; it thus refines the selection and presentation of the important sentences within the final summary

## III. SYSTEM ARCHITECTURE

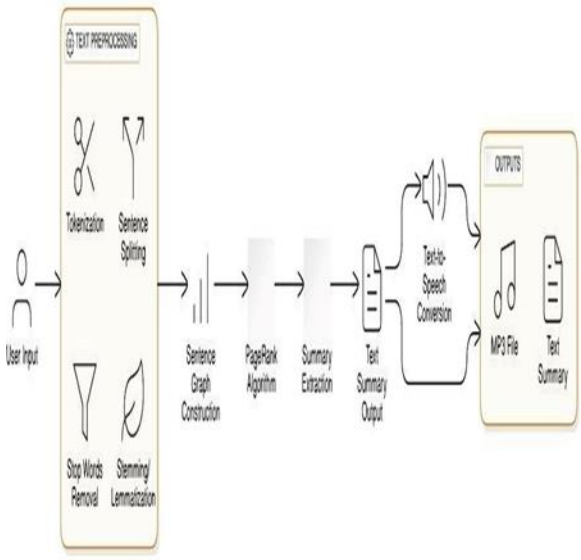


*Figure 1:* System Architecture

1. User Input: When a user feeds a text document into the system, such as a news article or report, then the

system starts working. The user can also feed the text to the system by uploading it directly.
2. Preprocessing: The procedure cleans the text for summarization. Sentence splitting is performed to analyze text at an individual level by splitting it into sentences. Stop words are removed: These are simply defined as common words that can be excluded without adding any meaning, such as the, and, or is.
3. Page Rank Algorithm: The Page Rank algorithm assumes that within the document, the more important sentences carry a higher score. Sentences that lead to many other important sentences are considered important in their own right. These ranks are then iteratively updated on such sentences until saturation.
4. Summary Extraction: That would mean the sentences that get selected for thesummary are those that get the highest ranks. The system selects the maximum ranked sentences from the text that represent key ideas from the original document. Then the summary follows a sequence logically ordered for a good flow.
5. Text and Audio Output: The last summary is provided in both text and audio form. The text is displayed on a screen, and it is open for saving in several other.

## IV. METHODOLOGY

A text summarization system designed by our organization uses PageRank algorithm, NLTK library, and Google Text-to-Speech (GTTS) for summarizing text and giving its output in audio. The methodology is a structured process to generate summaries and present them.

*PageRank Algorithm:*
This is a graph-based ranking approach that was originally proposed by Google to rank web pages. In our system, however, we use it to provide rankings of sentences in the document based on their importance. Steps of PageRank algorithm: 1. Constructs the sentence similarity graph - nodes are sentences; edges represent similarities between those sentences. 2. Edge weight calculations - Degree of similarity on between sentences is calculated by the cosine similarity. 3. Graph-based ranking - The PageRank algorithm assigns an importance score to each sentence. 4. Selecting sentences from the highest ranked sentences to form the final summary.

*Natural Language Toolkit (NLTK):*
NLTK means Natural Language Toolkit, which is a strong Python library used for text processing and analysis. It has some features of text preprocessing, like tokenization, stopword removal, stemming, and lemmatization. In the case of this system, NLTK helps in the processing of sentence segmentation, part-of-speech tagging, and Named Entity Recognition (NER) for better results in summarization. NLTK Word2Vec embeddings are also used to gain enhancements in representing the sentence for improving the PageRank-based ranking system accuracy.

*Google Text-to-Speech (GTTS):*
GTTS means Google Text-to-Speech. This Python library provides features to convert text into spoken words. The extracted text will be converted to speech using GTTS and saved as an MP3 fle after the summary is generated. It provides convenience and accessibility features such that a person can listen to the summary generated by the system instead of reading while he suffers from visual impairment. GTTS additionally provides a multi-language feature with the possibility to use several speech speeds, thus increasing the level of experience across various audiences. Basically, this orderly system allows for the elimination of enormous volumes of textual data provision of interactive audio output for user accessibility at the end point.

## V. CONCLUSION

The text tells how PageRank is applied to summarization in distilling lengthy news articles to a comprehensible state. Important sentences are selected based on their linkages among themselves for the minimization of the core ideas and the elimination of other trivial details. The essence of PageRank lies in ranking the sentences according to their importance, and as a result, generating a brief and informative summary. Such a procedure will be suitably employed for summarizing various document types, such as news articles, research papers, and blog posts. However, the system can be further enriched by fusing advances in artifcial

intelligence technologies, like machine learning and natural language processing (NLP).

REFERENCES


[1] A Comparative Study of Opinion Summarization Technique, Prof. Surbhi Bhatia, IEEE, https://doi.org/10.1109/TCSS.2020.3033810

[2] Summarization and Simplification of Medical Articles using Natural Language Processing, Shashank Patel, RuchaNargunde, ShobhitVerma,Dr. SudhirDhage,IEEE,and https://doi.org/10.1109/ICCCNT54827.2022.998449

[3] A Graph-to-Sequence Learning Framework for Summarizing Opinionated Texts, Penghui Wei, Jiahao Zhao, and Wenji Mao, IEEE, https://doi.org/10.1109/TASLP.2021.\

[4] Improving Unsupervised Extractive Summarization by Jointly Modeling Facet and Redundancy,Prof. Xinnian Liang , Jing Li, Shuangzhi Wu, Mu Li, and Zhoujun , IEEE,https://doi.org/10.1109/TASLP.2021.3138673

[5] Extractive Arabic Text Summarization Using PageRank and Word EEmbedding, GhadirAlselwiTŭgrul Ta , scı1, Arabian Journal for Science and Engineering (2024),https://doi.org/10.1007/s13369-024-08890-1

[6] A REVIEW PAPER ON TEXT-TO-SPEECH CONVERTOR", Sneha Tamboli1, Pratiksha Raut1, LavkushSategaonkar1, Anjali Atram1, Shubham Kawane1, Prof. V. K. Barbudhe2, International Journal ofResearch Publication and Reviews, Vol 3, no 5, pp 3807-3810, May 2022.

[7] "Jointly Learning Topics in Sentence Embedding for Document Summarization", Yang Gao ,Member ,YueXu , Heyan Huang , Qian Liu, Linjing Wei, and Luyang Liu, IEEE,https://doi.org/10.1109/TKDE.2019.2892430

[8] "A Joint Sentence Scoring and Selection Framework for Neural Extractive Document Summarization"", QingyuZhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and TiejunZhao,IEEE,https://doi.org/10.1109/TASLP.2020.2964427

[9] "Extractive Text Summarization Using Sentence Ranking", J.N.Madhuri , Ganesh Kumar.R, IEEE, https://doi.org/10.1109/IconDSC.2019.8817040