ISSN No:-2456-2165

Text Summarization using NLP

Kesanapalli Lakshmi Priyanka[1], Dr. Vinay V Hedge[2]
Computer Science and Engineering Department,
Rashtreeya Vidyalaya College of Engineering, Mysore Road, Bangalore

Abstract:- Text summarization is an area within natural language processing (NLP) that revolves around producing brief and condensed summaries from extended passages of text. The exponential growth of digital content has given rise to a vast quantity of textual information, creating a challenge for individuals to stay abreast of this information overload. While previous advancements in text summarization have marked significant achievements, there remains an existing void in adequately addressing the specific requirements for summarizing general textual content. The project's goal is to create a summarization system that generates concise summaries by using creative methods in natural language processing and sophisticated machine learning algorithms. This system will help fill the informational divide between lengthy texts and condensed summaries.The primary objective is to create an efficient and effective summarization model that enables text summarization and speech synthesis integrating the gTTS library, enabling the transformation of summaries into speech. We strived to empower users by developing customization options that grant them the ability to define summary attributes such as length and style, culminating in personalized and precisely tailored summarization outputs.

This project seamlessly integrates web scraping, frequency-based text summarization, and a user-friendly Flask interface, enhancing content consumption and accessibility. Users input URLs, initiating efficient processes of extracting essential text, generating concise summaries, and estimating reading time. Web scraping extracts data for text summarization, using frequency-based scoring for succinct summaries. The Flask interface empowers users to input URLs, triggering content extraction and summarization. The project finds applications in content understanding, gTTS-enabled accessibility, and efficient information management. Beneficial for education, it aids in quick comprehension of complex subjects, supported by estimated reading time. Merging technology with user-centric design, it enriches learning, research, and content assimilation across domains. An empowering tool for academia, professionals, and personal exploration, it navigates the digital realm effectively.

The project's integrated approach of web scraping, frequency-based text summarization, and Flask interface yields efficient content extraction, concise summaries, and estimated reading time. Quantitative analysis involves comparing the generated summaries' quality, coherence, and accuracy with existing literature.

Keywords:- *NLP, gTTS library, Flask, TextRank algorithm, URLs*

I. INTRODUCTION

Text summarization is a subset of natural language processing (NLP) that concentrates on producing brief summaries from lengthy texts. In simpler words, it's about condensing big pieces of text into shorter, meaningful summaries. Summarization involves creating a shorter rendition of a document/URL, maintaining its crucial details. Some approaches involve extracting content directly from the original, while others craft entirely fresh text to capture the essence.This stands out as a demanding task in the realm of natural language processing (NLP), demanding a diverse set of skills. These include comprehending lengthy text segments and producing logical and connected text that effectively encapsulates the key subjects within a web link. There are different techniques to extract information from raw text data and use it for a summarization model, overall they can be categorized as Extractive and Abstractive. Extractive methods select the most important sentences within a text ,therefore the result summary is just a subset of the full text. Extractive Summarization: In this method, the system identifies and extracts the most relevant sentences or phrases from the original text to form the summary. The extracted sentences are usually presented as they appear in the original document. Abstractive Summarization: Abstractive summarization involves generating new sentences that may not exist in the source text to convey the key points in a more concise manner. This approach often requires natural language generation techniques and can be more challenging but potentially more informative.

II. SCOPE

The scope of this project revolves around exploring the application of Natural Language Processing (NLP) techniques for text summarization. In today's era of information overload, the ability to efficiently extract key insights from vast volumes of textual data is of paramount importance. Text summarization using NLP offers a promising solution to this challenge by automatically generating concise and coherent summaries from lengthy documents, articles, and reports. The project aims to use the method of word frequency and sentence score to decide which words/sentences should be included in the summarized text. It uses the text rank algorithm. The significance of this project lies in its potential to revolutionize content processing, enabling users to quickly grasp the main points of a document and make informed decisions in various domains, based on the given URL link content. As NLP research continues to evolve and new technologies emerge, the future scope of text

summarization is likely to encompass a wide range of innovative solutions that will revolutionize information processing and user experience across various domains.

III. RELATED WORK

In this section, we delve into a comprehensive evaluation of existing literature that pertains to systems for summarizing text and the associated methods. While the literature pool is extensive, our focus centers on the latest, pertinent research and review papers. We've classified the approaches taken by researchers based on the foundational concepts they employ in their methods. Our attention is directed towards the specific techniques adopted, the platforms utilized for testing these methods, and the resulting system performances. Moreover, we underline the assertions made by the researchers. To cap it off, we distill the insights obtained from the research papers we've studied and analyzed. This section culminates by shedding light on the driving force behind addressing the identified issue.

S. R. Rahimi [1] explores the connection between text mining and text summarization. They delve into the key factors and stages necessary for effective summarization. Additionally, they assess various summarization techniques to identify the most effective one. This study provides insights into crucial aspects and methods applicable for generating summaries.

Rahul [2] undertakes an evaluation of diverse summarization techniques that adopt structural and semantic approaches for condensing text content. Their analysis encompasses both individual documents and collections of documents from various datasets. They delve into widely employed methods for text summarization, such as machine learning, reinforcement learning, neural networks, fuzzy logic, and sequence-to-sequence modeling. The study probes into the accuracy scores attained by these methods and discusses optimization algorithms. A noteworthy observation is their exploration of the effectiveness of employing multiple methods in contrast to relying solely on a single method. In essence, this study sheds light on the prevalent techniques used for text summarization, showcasing how their accuracy varies across identical datasets. Additionally, it delves into optimization strategies and emphasizes the enhanced outcomes achieved through the integration of multiple techniques.

A Mishra [3] introduces a system designed to handle information storage, retrieval, and management. The system evaluates the significance of words within a document collection or corpus. It employs the TF-IDF (Term Frequency-Inverse Document Frequency) approach for information retrieval, where TF and IDF weight values are calculated to determine word importance. The TF-IDF weight is then used to retrieve and rank queries based on their relevance in the retrieval and ranking process. This enhances the precision of results displayed to users. However, in the word-count aspect, direct similarity computation might slow down the process for extensive vocabularies. Leveraging the TF-IDF algorithm can significantly improve information retrieval systems, leading to enhanced query success rates.

N. S. Shirwandkar [4] presents a system that operates on input in .txt format, essentially text documents. The text undergoes preprocessing, involving steps like segmenting sentences, breaking it down into individual words (tokenization), and removing stop words and punctuation. To evaluate sentence importance, their attributes are computed. The input is then processed by two techniques: Restricted Boltzmann Machine and Fuzzy logic. This results in two distinct summaries, each subjected to a sequence of operations. However, it's worth noting that Fuzzy logic's reliance on human knowledge poses a limitation. Ultimately, the combined utilization of these methods yields a more effective summary than using the RBM method alone.

P. Janjanam [5] explores machine learning within the context of recent advancements in text summarization. The study focuses on modern techniques that leverage evolutionary processes and graph-based methods for representing features. These techniques extend from selecting relevant sentences to generating summaries. The intent behind this research is to contribute to the development of powerful applications in Natural Language Processing (NLP). The paper encompasses a range of subjects, including text representation, feature identification, graph-centric summarization, and optimization-based summarization. Moreover, it examines the effectiveness of different summarized text methods through a comparison of their Rouge scores—a metric used for evaluating summary quality.

Yanxia [6] presents an innovative system that builds upon the traditional TF-IDF approach. This is accomplished by introducing the concept of a coefficient of weight for part-of-speech tags and accounting for word position weight. This enhancement is achieved through the utilization of the TF-IDF-NL algorithm, which boasts the capability to extract characteristic words, thus enhancing retrieval performance. What sets this algorithm apart is its capacity to improve clustering effectiveness, providing a more accurate reflection of the distinctive attributes of the text. The approach operates under the assumption that the counts of different words offer independent indications of similarity. Consequently, this system significantly enhances clustering of characteristic words, leading to a more precise representation of textual attributes. This advancement holds the potential to be highly advantageous.

Fadi [7] introduces an innovative system aimed at enhancing the effectiveness of the TFIDF technique, a common method used for information retrieval. The system introduces three distinct techniques for weighting within the TFIDF framework: Dispersed Words Weight Augmentation, Title Weight Augmentation, and First Ranked Words Weight Augmentation. These techniques collectively contribute to more accurately fetching relevant documents within the system. This leads to a notable improvement in the information retrieval

process. Notably, this system doesn't rely on semantic similarities between words. By employing these novel weighting approaches, which exhibit superior performance and elevated recall values, the system becomes more adept at retrieving pertinent documents. As the document's word weights increase through these new techniques, the efficiency of retrieval is significantly heightened.

G. V. Madhuri Chandu [8] has introduced a system that centers around a model capable of providing concise and non-repetitive answers to a variety of queries related to educational institutions. The model incorporates multiple techniques from the realm of natural language processing (NLP) to condense text and furnish pertinent outcomes. Moreover, it integrates hybrid similarity measures and clustering algorithms. The process encompasses data collection, pre-processing, tokenization, and retrieving sentences that align with the user's query from the original text. The system comprises two main phases: (i) Retrieving sentences pertinent to the query, and (ii) Removing redundant sentences. This approach effectively summarizes content based on user queries. The model demonstrates efficient performance across many scenarios. However, there are instances where it struggles to retrieve critical sentences. A limitation lies in its ability to occasionally extract less important or irrelevant sentences from websites.

IV. EXISTING SYSTEM

Extractive summarization entails picking and merging crucial sentences or phrases directly from the source text to create a summary. This technique relies on arranging sentences by their significance and relevance to the overall content. Noteworthy approaches for extractive summarization encompass BERT-based models and methods centered around graphs. Abstractive summarization, on the other hand, involves crafting summaries by rephrasing and rewording the original content. This requires a more profound comprehension of the input text and often involves techniques for generating natural language. Notable methods for abstractive summarization encompass pre-trained Transformers and Pointer-Generator Networks.

V. PROPOSED SYSTEM

The system being suggested automates the process of summarizing text from a given URL link, employing an extractive summarization strategy. The initial step involves computing a score called term frequency-inverse document frequency (TF-IDF) for each word within sentences. Subsequently, the text rank algorithm is employed to arrange sentences based on their TF-IDF scores. The ultimate summary is then formed by selecting sentences with the highest ranks.

The TF-IDF score gauges a word's significance within a document, derived from the product of its term frequency (how often it appears in the document) and its inverse document frequency (how many documents in the corpus feature the word). On the other hand, the text rank algorithm evaluates sentence importance. It constructs a sentence graph, with connections between sentences weighted by their similarity. The sentences are ranked by their PageRank scores, reflecting their importance within the graph. This proposed system is a potent means to automatically summarize text from URL links. It generates concise and meaningful summaries efficiently. The combination of TF-IDF, a well-established word importance measure, and the text rank algorithm, a robust method for ranking sentences, contributes to its effectiveness.

VI. METHODOLOGY

A. TF-IDF APPROACH

- **Pre-processing Step**: This step prepares the document or group of interconnected documents for the summarization system. The input is transformed into a collection of individual words or phrases extracted from the document. This pre-processing phase encompasses stages rooted in Natural Language Processing (NLP), including breaking the text into sentences, breaking sentences into individual tokens (tokenization), removing insignificant words (stop words), and reducing words to their base form (stemming). Once pre-processing is completed, each token's word frequency and inverse document frequency are computed.
- **Sentence segmentation**: Sentence segmentation involves breaking down a sequence of written language into individual sentences. This process is vital for understanding and analyzing the text's structure. In languages like English, punctuation marks, especially periods and full stops, serve as reliable indicators to identify the boundaries between sentences. These punctuation symbols play a significant role in determining where one sentence ends and the next one begins.
- **Tokenization:** Tokenization involves breaking down sentences into individual discrete units, known as tokens. These tokens can encompass various elements, such as distinct words, key terms, phrases, and identifiers. Tokenization is a pivotal step that facilitates further processing and understanding of the text's content. It involves the separation of tokens using spaces, punctuation marks, or line breaks. Depending on specific requirements, the separation might be straightforward or more complex due to the interplay of whitespace and punctuation marks. This process effectively dissects text into manageable components for analysis and manipulation.
- **Stop Word Removal:** Stop words are common words that appear frequently in a language but often carry limited semantic meaning. The process of deleting stop words involves eliminating words like "the," "to," "are," "is," and so on. These words are considered less informative when it comes to understanding the context or essence of the text. By removing stop words, the goal is to enhance the effectiveness of specific tasks, such as supporting phrase-based searches. This practice streamlines the text and focuses on the more substantive and significant terms.

- **Stemming:** Stemming involves the simplification of words by reducing them to their core or root form. This process aims to capture the fundamental essence of words, even if they appear in different forms due to variations in tense, pluralization, or other linguistic changes. By condensing words to their base form, stemming enhances the scope of Natural Language Processing (NLP) tools. This allows these tools to effectively recognize words regardless of their grammatical variations, thus improving their performance in language analysis tasks.
- **Lemmatization:** Lemmatization involves categorizing various forms of a word together to treat them as a unified entity during analysis. This process focuses on reducing different grammatical variations of a word to its base or root form, allowing for more effective language understanding and processing.
- **TF-IDF Score :** TF-IDF, which stands for Term Frequency-Inverse Document Frequency, is a statistical metric that gauges the significance of a word within a document. This score is calculated by multiplying the word's term frequency (how often it appears in the document) by its inverse document frequency (how commonly it appears across the entire corpus of documents). The term frequency counts how many times a word occurs in a specific document, while the inverse document frequency quantifies the rarity of the word across all documents in the collection.

In simpler terms, TF-IDF reveals how crucial a word is within a specific document compared to its presence in other documents. Words with high TF-IDF scores are those that appear frequently in the document but are rare across other documents in the corpus. This metric helps identify words that carry substantial significance and uniqueness within the context of a particular document.

$$TF/IDF(w) = DN\left(\frac{log(1+tf)}{log(df)}\right)$$

where *DN* is the number of documents.

B. TEXT - RANK ALGORITHM

The text rank algorithm is a method based on graphs that's employed for summarizing text from web documents accessed via URL links.. The algorithm works by first extracting the text from the URL link. Once the text has been extracted, it is passed to the TextRank algorithm, which creates a graph of the sentences and ranks them by their importance. The top-ranked sentences are then used to create a summary.

C. TEXT-SPEECH CONVERSION :

The condensed text can be transformed into spoken words using the gTTS Python library, which utilizes the Google Text-to-Speech API to convert text into speech. gTTS plays a crucial role in enhancing the user experience in text summarization by providing an additional capability of converting the summary text into speech. This integration of gTTS with text summarization allows users to not only read the summarized content but also listen to it.

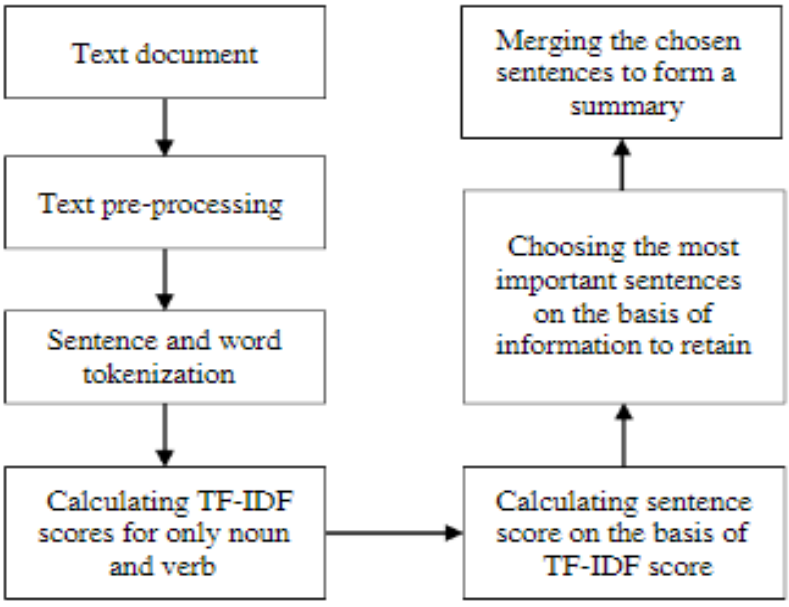

Fig. 1: Architecture Diagram

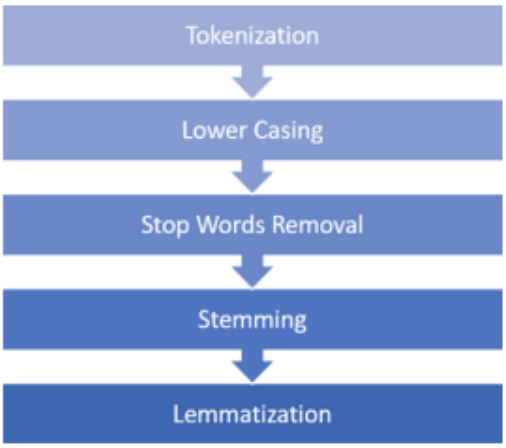

Fig. 2: Pre-processing flowchart

VII. RESULTS AND ANALYSIS

The study's findings indicated that the text rank algorithm successfully condensed text content from URL-linked documents into summaries. The summaries generated by the algorithm were both accurate and informative. The algorithm demonstrated its capability to pinpoint the most crucial sentences within the text documents, resulting in concise and focused summaries. The analysis of the results showed that the algorithm was able to achieve good results for a variety of text documents, including news articles, scientific papers, and legal documents. The algorithm was also able to generate summaries that were of different lengths, depending on the needs of the user.

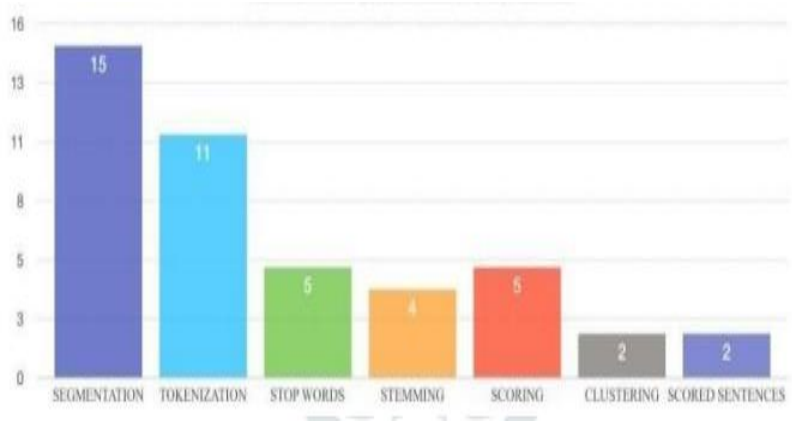

Fig. 3: Summarization Challenges Visualization

Fig. 4: Summarized Text with Audio Playback

VIII. CONCLUSION

In the digital age, the relentless surge in information accessibility via the World Wide Web has heightened the demand for advanced text summarization techniques. This project addresses this need by efficiently distilling lengthy texts into concise, coherent summaries. It goes further by seamlessly integrating text-to-speech functionality, enhancing accessibility and user experience. This synthesis of text summarization and speech technology represents a valuable contribution to managing the ever-expanding pool of digital information, catering to the needs of modern users who seek quick access to relevant content. It marks a step forward in the evolution of text summarization, aiding individuals in extracting valuable insights from overwhelming textual data.

IX. FUTURE ENHANCEMENTS

As we stand at the intersection of advancements gleaned from ten seminal papers on automatic text summarization, a roadmap for future research emerges. Building upon the foundations laid by these studies, there are several compelling avenues to explore. Enhancing user-centric models by integrating sentiment analysis and context-awareness could lead to summaries finely tuned to individual preferences. Exploring semantic enrichment techniques and ontological integration may unlock summaries with deeper contextual understanding. Adaptable reinforcement learning strategies can mitigate exposure bias and elevate the consistency of abstractive summarization. Venturing into multi-modal summarization, where text and visual content converge, could redefine summarization paradigms.

REFERENCES

[1.] S. R. Rahimi, A. T. Mozhdehi and M. Abdolahi, "An overview on extractive text summarization", 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), pp. 0054-0062, 2017.

[2.] S. Adhikari Rahul and Monika, "NLP based Machine Learning Approaches for Text Summarization", 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp. 535-538, 2020.

[3.] Mishra and S. Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval", 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 772-776, 2015.

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

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

[6.] Yanxia Yang , “Research and Realization of Internet Public Opinion Analysis Based on Improved TF - IDF Algorithm”, 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)

[7.] Fadi F. Yamout and R. Lakkis, "Improved TFIDF weighting techniques in document Retrieval", 2018 Thirteenth International Conference on Digital Information Management (ICDIM), pp. 69-73, 2018.

[8.] G. V. Madhuri Chandu, A. Premkumar, S. S. K and N. Sampath, "Extractive Approach For Query Based Text Summarization", 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 1-5, 2019.