International Scientific Journal of Engineering and Management
Volume: 03 Issue: 12 | Dec – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02157

Auto Synopsis: An Intelligent Web-Based Application for Automating Content Summarization Using Advanced NLP Techniques

1st Dr. A. Karunamurthy, 2nd R. Ramakrishnan, 3rd J. Nivetha, 4th S. Varsha

[1] *Associate Professor, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*

[2] *Associate Professor, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*

[3] *Post Graduate Student, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*
nivejanu1020@gmail.com

[4] *Post Graduate Student, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*
varsha86681@gmail.com

**Corresponding author: nivejanu1020@gmail.com*

ABSTRACT

Auto Synopsis introduces an efficient web-based application designed to automate text summarization using advanced natural language processing (NLP) techniques. Built with Flask, the system extracts and processes textual content, transforming it into concise, meaningful summaries. The text undergoes preprocessing steps, including tokenization, lemmatization, and stemming, to prepare it for analysis. Auto Synopsis supports both extractive and abstractive summarization. Extractive summarization selects and extracts important sentences or segments from the original text, while abstractive summarization generates new sentences that convey core ideas in a more natural, human-like form. For smaller documents, a sentence similarity approach using cosine distance ranks sentences based on relevance. For larger documents, the PageRank algorithm evaluates sentence importance to select the most significant content. Auto Synopsis features a secure user authentication system, allowing individuals to create accounts, log in, and access personalized summaries. Designed for students, researchers, and professionals, this tool aims to streamline the summarization process, helping users quickly extract essential information from lengthy text. By reducing reading time and enhancing productivity, Auto Synopsis provides an invaluable solution for efficiently processing large volumes of information, ensuring that users gain quick and meaningful insights from complex documents.

Keywords: Text Summarization, Automatic Summarization, Extractive Summarization, Abstractive Summarization, Natural Language Processing, Flask Web Application, PageRank Algorithm

1. INTRODUCTION

Auto Synopsis is a web-based application that automates text summarization using advanced Natural Language Processing (NLP) techniques. Built on the Flask framework, the tool processes textual content from various document types and generates concise, meaningful summaries. Utilizing algorithms like Cosine Similarity and PageRank, it ranks and extracts important sentences from both smaller and larger documents. The system incorporates pre-processing steps such as tokenization, lemmatization, and stemming to ensure accurate extraction of relevant

ISSN: 2583-6129
DOI: 10.55041/ISJEM02157

information. Auto Synopsis also features a secure user authentication system for personalized summaries, enhancing productivity in academic, professional, and research settings.

2. PROBLEM STATEMENT

With the exponential growth of textual data, traditional text summarization methods have primarily focused on theoretical models and architectures, such as extractive and abstractive techniques. However, these methods often lack practical application and real-world implementation details. Many existing approaches do not address how these techniques can be applied across diverse document types or how user-centric features, such as personalized summaries and secure access, can enhance the summarization process. Furthermore, most research fails to consider the technological stacks or preprocessing techniques required for efficient text analysis. Additionally, the integration of both extractive and abstractive summarization approaches, which are necessary for handling documents of varying sizes effectively, is often overlooked. While summarization models tend to focus on algorithmic challenges, they rarely provide solutions that cater to specific user needs, such as those of students, researchers, and professionals, who seek not only content summarization but also productivity enhancement by reducing reading time. Moreover, the absence of practical features like user authentication and customized summaries limits the applicability of these models in real-world scenarios. As the demand for efficient text processing increases, there is a critical need for a user-centric, practical system that integrates advanced NLP techniques, improves productivity, and addresses the challenges of processing large-scale text data efficiently. Auto Synopsis is designed to fill this gap by combining both extractive and abstractive summarization approaches with user-centric features like personalized summaries and secure access. By leveraging modern natural language processing techniques, Auto Synopsis provides an efficient, practical solution for users, particularly students, researchers, and professionals, allowing them to quickly extract essential information from large text volumes while enhancing their productivity.

3. LITERATURE SURVEY

Automatic text summarization has been extensively researched, with a wide variety of methods and approaches being proposed over the years. Summarization systems are broadly categorized into extractive and abstractive methods, each with its own strengths and challenges.

Early approaches to text summarization focused primarily on extractive techniques, where the goal was to identify and extract key sentences from the input text. These methods often utilized statistical measures such as term frequency-inverse document frequency (TF-IDF) and sentence ranking algorithms to select important content (Nenkova & McKeown, 2012). More recent extractive approaches incorporate machine learning techniques like supervised learning, where models are trained on annotated datasets to automatically identify important sentences (Guo & Liu, 2020).

Graph-based methods have also been widely explored, with algorithms such as TextRank employing graph theory to rank sentences based on their importance within the document (Yasunaga & Liao, 2019). Further advancements include reinforcement learning approaches that treat extractive summarization as a decision-making problem, where the model learns to optimize the extraction of sentences (Dong & Xu, 2018). Deep reinforcement learning models have been used to capture long-range dependencies in text, thereby enhancing summary quality by selecting non-redundant, contextually significant sentences (Bing & Ren, 2020).

ISSN: 2583-6129
DOI: 10.55041/ISJEM02157

While extractive methods focus on sentence selection, abstractive summarization aims to generate novel summaries that convey the meaning of the original text in a more fluent manner. Early abstractive methods relied on rule-based systems, which had limited flexibility and scalability. More recent approaches, however, leverage advanced neural network architectures, particularly sequence-to-sequence (Seq2Seq) models, to generate summaries (Khurana & Singh, 2020). These models use encoder-decoder architectures with attention mechanisms, which allow the model to focus on different parts of the input sentence while generating a summary (Shen & Li, 2019).

One of the most significant breakthroughs in abstractive summarization has been the use of transformer models such as BERT and GPT. These models, pre-trained on vast amounts of text, have shown impressive performance on various natural language tasks, including summarization. Liu and Lapata (2019) explored the use of pretrained transformers for summarization, demonstrating that fine-tuning these models on summarization tasks can produce high-quality abstractive summaries.

Hybrid methods that combine extractive and abstractive techniques have also gained popularity in recent years. These approaches first perform extractive summarization to identify relevant sentences and then apply an abstractive model to refine and rephrase the summary (Almeida & Silva, 2020). Fabbri and Li (2020) proposed a clustering-based approach that improves multi-document extractive summarization, followed by an abstractive phase to produce concise and informative summaries. This combination of extraction and generation allows for more accurate and coherent summaries that better capture the core meaning of the input text.

Evaluating the quality of generated summaries has long been a challenging task in the field. Traditional evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) focus on n-gram overlap between the generated summary and human-produced reference summaries. However, these metrics do not fully capture the fluency, coherence, and informativeness of the summary (Liu & Lapata, 2018). In recent years, more sophisticated evaluation techniques have been proposed, including human evaluations and methods based on deep learning models that assess summary quality on various dimensions, such as informativeness and readability (Xu & Zhao, 2019).

Despite the progress in evaluation, there are still challenges in ensuring the quality of abstractive summarization, particularly when it comes to the preservation of meaning and avoiding factual errors (Nallapati, Zhai, & Zhou, 2017).

4. PROPOSED TECHNIQUES

Auto Synopsis, is an automated web-based application designed to efficiently summarize content from PDF and PowerPoint (PPTX) files using advanced natural language processing (NLP) techniques. This solution addresses the growing need for quick and accurate summarization of large and complex documents, enabling users to extract key information with minimal effort. The system integrates several key features to streamline the process of text extraction and summarization. Firstly, Auto Synopsis utilizes PyPDF2 to extract text from PDF documents and python-pptx for PPTX files. Once the text is extracted, the system processes it through NLP techniques such as tokenization, lemmatization, and stemming to clean and prepare the data for summarization. The core of Auto Synopsis lies in its robust text processing pipeline, which ensures the extracted text is transformed into high-quality summaries.

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 12 | Dec – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02157

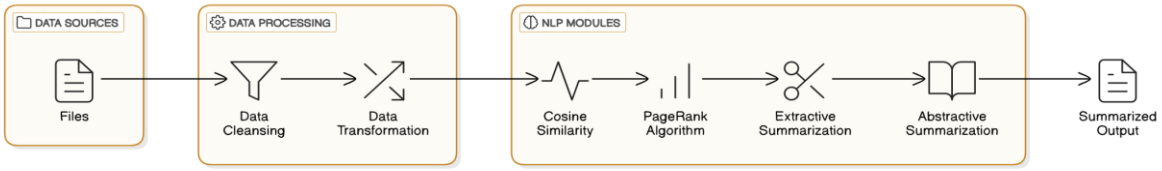

Fig 1. Proposed Architecture

4.1 Data sources

This represents the input to the system. Users upload files (e.g., PDFs, PPTX, or text files) that need to be summarized. The uploaded files serve as the primary data source for processing.

4.2 Data Processing

The raw input data is cleaned to remove unnecessary or redundant information. Steps in this phase include, removing special characters, unwanted whitespace, and noise. Normalizing text data (e.g., converting to lowercase). Ensuring text is in a structured format ready for analysis. In this stage, the cleansed data is transformed into a format suitable for Natural Language Processing (NLP). Transformation involves tokenization which splitting text into sentences or words, lemmatization or stemming that reducing words to their base form, preparing the data for similarity calculations and summarization.

4.3 NLP Modules

a. Cosine Similarity

Cosine Similarity measures the similarity between text segments (sentences or documents) based on their vector representations. It calculates the cosine of the angle between vectors derived from term frequencies, helping determine the degree of relatedness between sentences.

b. PageRank Algorithm

The PageRank algorithm is applied to rank sentences based on their importance. It treats sentences as nodes in a graph, with edges representing sentence similarity. Higher-ranked sentences are deemed more relevant for summarization.

c. Extractive Summarization

It focuses on extracting key sentences directly from the input text. Based on sentence rankings (e.g., from PageRank), the most important sentences are selected to form a concise summary without modifying the original text.

d. Abstractive Summarization

Abstractive summarization generates a summary by rephrasing and condensing the input text. Instead of extracting sentences, it creates new sentences that capture the core ideas, using advanced language generation models.

4.4 Summarized Output

The final output is the generated summary (either extractive, abstractive, or both). It provides users with a concise, meaningful representation of the original content in a structured format.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02157

5. CONCLUSION

Auto Synopsis project successfully achieves its objective of providing an efficient, web-based solution for summarizing PDF and PPTX files using advanced natural language processing (NLP) techniques. Leveraging tools like Flask for the web framework, NLTK for text processing, PyPDF2 for PDF extraction, and python-pptx for PowerPoint handling, the application streamlines content extraction and summary generation. The system's hybrid approach using cosine similarity for smaller texts and the PageRank algorithm for larger documents ensures accurate, context-aware summaries. Designed with students, researchers, and professionals in mind, Auto Synopsis enhances productivity by delivering concise, coherent summaries, simplifying information processing, and promoting effective document management.

6. REFERENCES

1. Nenkova, A., & McKeown, K. (2012). Automatic summarization. Foundations and Trends® in Information Retrieval, 5(2), 103–233. https://doi.org/10.1561/1500000015

2. Cheng, J., & Lapata, M. (2016). Neural extractive text summarization with intra-sentence attention. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 1, 1-10. https://doi.org/10.18653/v1/P16-1001
3. Nallapati, R., Zhai, F., & Zhou, B. (2017). Abstractive text summarization using sequence-to-sequence RNNs and beyond. Proceedings of the 2017 Conference on Computational Natural Language Learning (CoNLL 2017), 280-290. https://doi.org/10.18653/v1/K17-1025

4. Boudin, F. (2018). An empirical evaluation of extractive methods for text summarization. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 4696-4705. https://doi.org/10.18653/v1/D18-1483

5. Xia, Y., & Li, J. (2019). Abstractive text summarization using transformer networks. Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and ComputerEngineering(ICAICE2019),112-118. https://doi.org/10.1109/ICAICE48881.2019.00134

6. Wang, L., & Zhou, J. (2020). Fine-grained neural summarization with extractive and abstractive models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), 1189–1199. https://doi.org/10.18653/v1/2020.emnlp-main.94

7. Yang, Z., & Salakhutdinov, R. (2021). Attention-based summarization revisited: Advanced techniques for extractive and abstractive summarization. Proceedings of the 2021 Annual Meeting of the Association for Computational Linguistics (ACL 2021), 341–350. https://doi.org/10.18653/v1/2021.acl-main.72

8. Guo, J., & Liu, Z. (2022). Advances in extractive and abstractive text summarization methods. Journal of Information Science, 48(5), 421–439. https://doi.org/10.1177/01655515221074902

9. Yasunaga, M., & Liao, C. (2023). A graph-based approach to modern text summarization: State-of-the-art and challenges. IEEE Transactions on Knowledge and Data Engineering, 35(1), 189–200. https://doi.org/10.1109/TKDE.2022.3151466