Rajiv Gandhi University of Knowledge Technologies-Andhra Pradesh

RK Valley Institute

(Constituted under the A.P Govt Act 18 2008 and recognized as per Section 2(f,12(B) of UGC Act,1956) Accredited by '**NAAC**' with '**B+**' Grade

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NEWS ARTICLE AND PDF SUMMARIZER

Under the Guidance of

V.Sravani

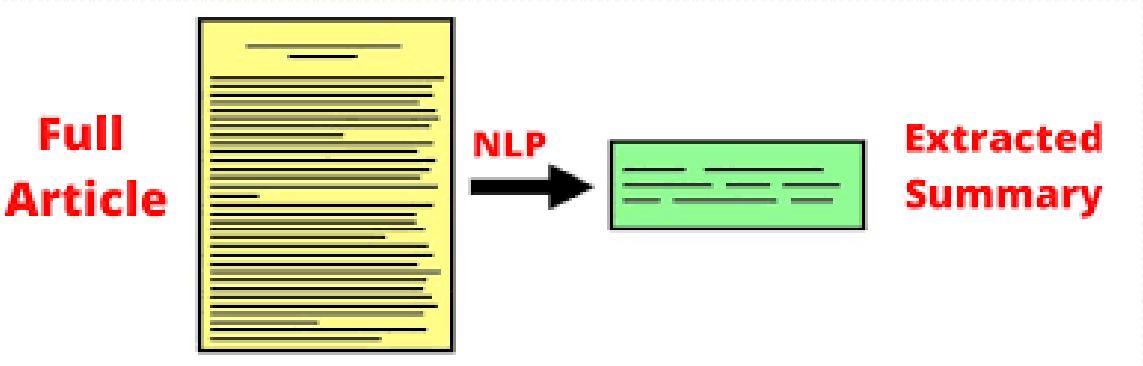
Team Members:

R190461: M.Nirosha

R191097: G.Pavani

Abstract

The News Summarizer project addresses this challenge by developing a tool that efficiently extracts and summarizes news articles,allowing users to quickly grasp the key points without needing to read entire articles.



- **User-Friendly Interface:**
  The application is designed with user convenience in mind,featuring a user-friendly interface built using Python's Tkinter library.

- **Flexible Input Formats:**
  The tool accepts both URLs of online news articles and PDF files containing news content, offering flexibility in how users can input the information to be summarized.

- **Error Handling and Reliability:**
  Error handling mechanisms manage issues like network errors, invalid URLs, and unsupported PDF formats. This ensures a smooth and reliable user experience.

- **Future Enhancements:**
  Future plans include integrating advanced NLP techniques and machine learning models for improved summary accuracy and relevance.

- **Advanced NLP Integration:**

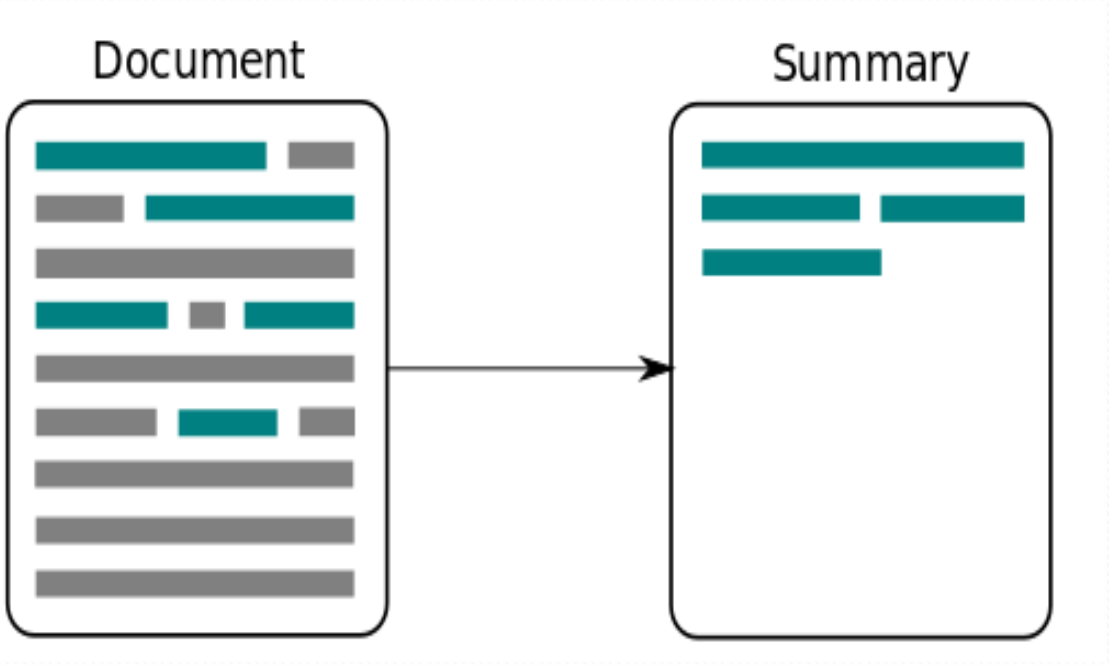
  The tool uses the newspaper3k library to download and parse content and employs TextBlob for NLP tasks like tokenization and sentiment analysis. This provides insights into the overall tone of the article.

- **Standout Features:**

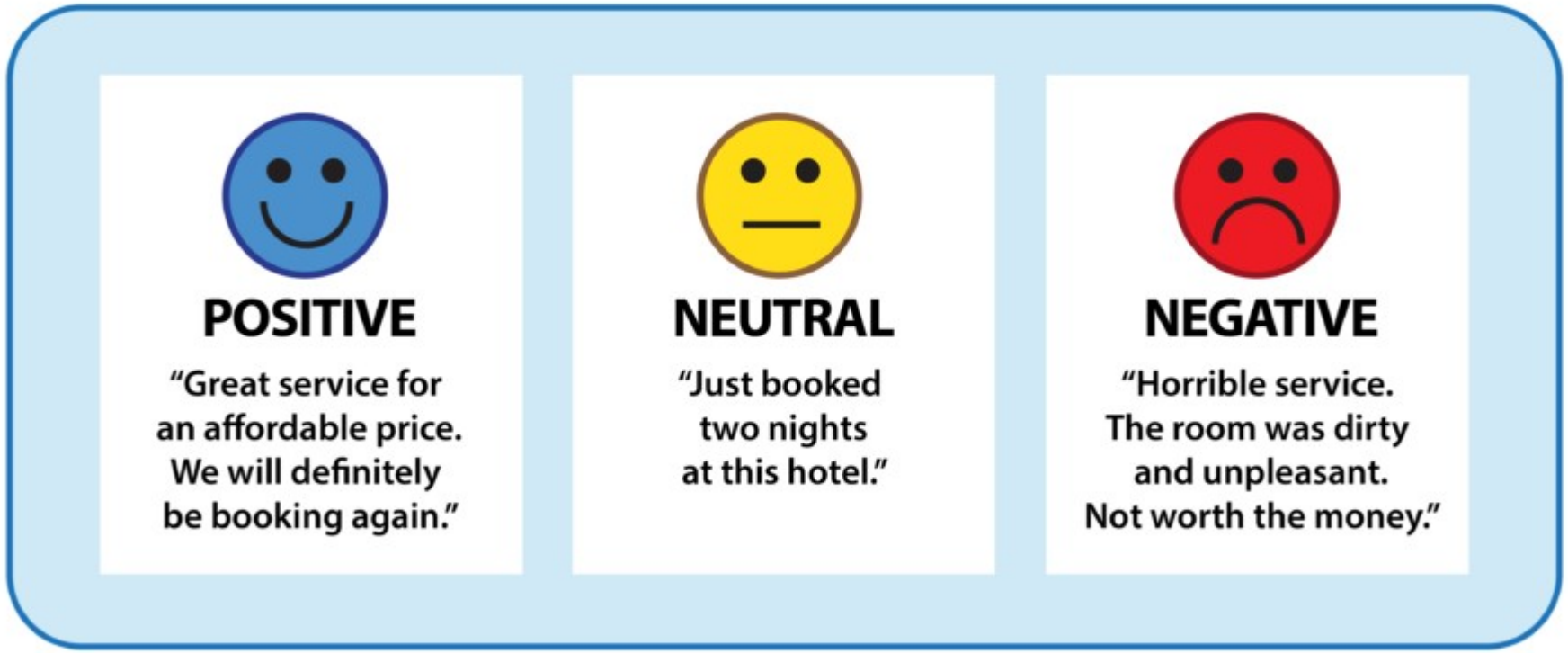
  The News Summarizer can save summarized content in PDF format, useful for archiving or sharing concise news versions. Summarization results and sentiment analysis are displayed directly within the Tkinter interface.

INTRODUCTION

- The News Summarizer project efficiently extracts and summarize news articles, helping readers quickly grasp key points without reading entire articles.
- This tackles the challenge of overwhelming daily news content making news consumption more manageable and less time-consuming for modern readers.



SENTIMENT ANALYSIS



Given text, sentiment analysis classifies its emotional quality.

- The primary objective of this project is to develop a robust and efficient News Summarizer Application that enhances the way users consume news and information.

- This application will leverage advanced natural language processing techniques to automatically generate concise.

- The primary functionalities include extracting,summarizing content from online articles and PDFs, and performing sentiment analysis determine whether the tone of the text is positive, negative, or neutral.

- This project contributes significantly to the field of information technology by enhancing the efficiency of content consumption and comprehension.
- It offers a comprehensive solution for summarizing lengthy articles and academic papers.
- The dual inputcapability, supporting both URLs and PDFs, increases the tool's versatility and applicability across different domains.

LITERATURE

This literature review describes different summarization techniques used for text summarization.

[1] Dipanjan Das Andre F.T. Martins (November 21, 2007)
This survey emphasizes extractive approaches to summarization using statistical methods.

[2] Archana AB, Sunitha. C (2013)
Archana AB, Sunitha. C describes comparative study on four different approaches to automatic text summarization.

[3] Saranyamol C S and Sindhu L (2014)
In this paper the author describes about the various techniques used in automatic text summarization which are extractive text summarization and abstractive text summarization respectively.

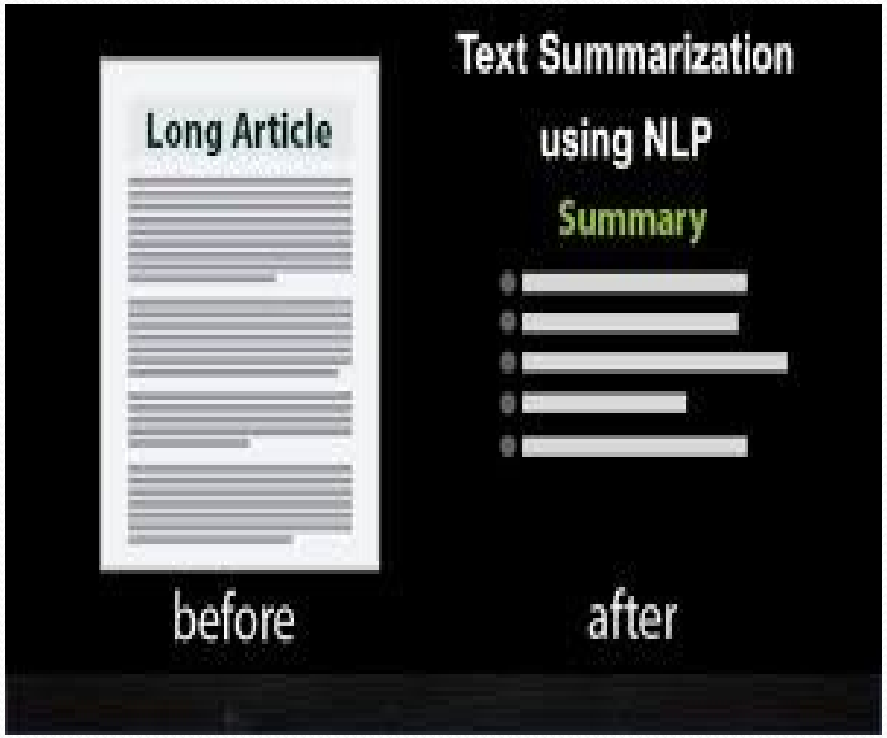
[4] K. Vimal Kumar, DivakarYadav(2015)
This paper mainly laid emphasis most importantly on the Hindi text summarization. The author had proposed a system which can generate the summary with 85 % accuracy.

[5] Richa Sharma, Prachi Sharma (April 2016)
This survey paper gives the details on extractive text summarization features and its methods. The extractive text summarization is a process of selecting important sentences from the document and including those sentences

[6] Vishal Gupta and Gurpreet Singh Lehal, (August 2010.)
In this paper author describes the extractive summarization methods which comprises of two parts PreProcessing and Processing.

Motivation

- The rapid growth of digital content has led to overwhelming influx of information,making it challenging for individuals to keep up with news and updates.
- Traditional methods of news consumption are becoming increasingly impractical, as they require significant time.



- Furthermore, the integration of both web and PDF input capabilities expands the tool's utility, making it versatile for various use cases.

- The motivation also extends to the educational sector, where such a tool can support students in managing their reading loads more effectively, thereby enhancing their learning experience.

- Overall, the project seeks to contribute to the digital information ecosystem by providing a reliable and efficient means of content summarization and sentiment analysis.

Contribution

- This project contributes significantly to the field of information technology by enhancing the efficiency of content consumption and comprehension.

- It offers a comprehensive solution for summarizing lengthy articles and academic papers, which is invaluable for researchers, students, and professionals who need to process vast amounts of information quickly.

Libraries Used

This project utilizes several key libraries, each serving a unique purpose to provide comprehensive functionality for the news summarizer application.

1.Tkinter
2.NLTK (Natural Language Toolkit)
3.TextBlob
4.Newspaper
5.Fitz(PyMuPDF)

Newspaper3 Library:

- The newspaper3k library is a powerful tool designed for web scraping and article extraction. It simplifies the process of downloading and parsing articles from the web, making it an essential component of our project.

Purpose: The primary purpose of the newspaper3k library is to provide a clean and structured way to extract articles from various websites.

Functionality: The library offers several functionalities, including:

- Downloading articles: It fetches the content from the provided URL.
- Parsing articles: It processes the fetched content to extract meaningful text, authorship information, publication date, and other metadata.
- Natural Language Processing (NLP) tasks: It includes basic NLP tasks such as extracting keywords, summaries, and top image from the article.

PyMuPDF (Fitz):

Handling offline documents is an essential feature for a versatile summarizer. The PyMuPDF library, also known as Fitz, provides capabilities to work with PDF files, enabling the extraction of text from these documents.

Purpose: The PyMuPDF library is used to allow the application to read and extract text from PDF files,broadening the range of input sources beyond just URLs.

Functionality: This library offers the following functionalities:

- Opening PDF files: It can open and read PDF documents, which are common in many professional and academic settings.

- Extracting text: It allows the extraction of text content from each page of the PDF, ensuring that the summarizer can process the document's content.

Usage: When a user selects a PDF file, the PyMuPDF library opens the document and extracts text from it.

NLTK(Natural Language Toolkit):

Text preprocessing is a critical step in any natural language processing (NLP) task. The nltk library, a leading platform for building Python programs to work with human language data, is employed for thispurpose in the News Summarizer project.

Purpose: The nltk library is used for tokenizing text into sentences and words, which is essential for preparing the text for summarization and other NLP tasks.

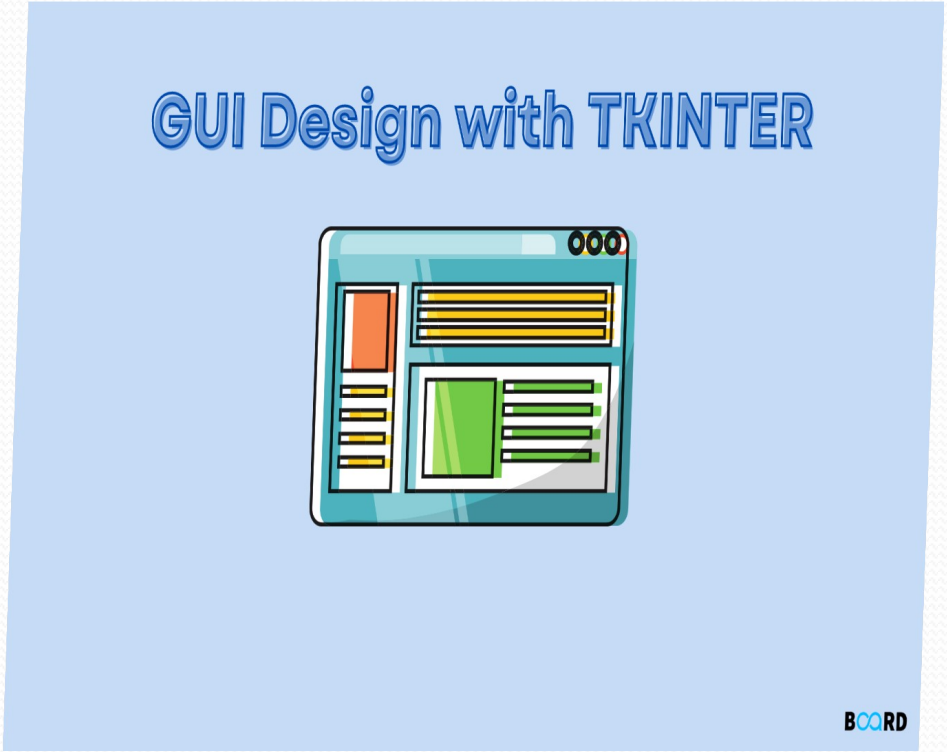
Functionality: The library provides a wide range of functionalities, including:

• Tokenization: Splitting text into sentences and words.
• Stemming and Lemmatization: Reducing words to their root forms.

Usage: In the News Summarizer project, nltk is primarily used for tokenizing the extracted text into sentences.

Tkinter:

Tkinter is a standard GUI (Graphical User Interface) library in Python that allows you to create windows, dialogs, buttons, menus, and other GUI elements for your Python applications.



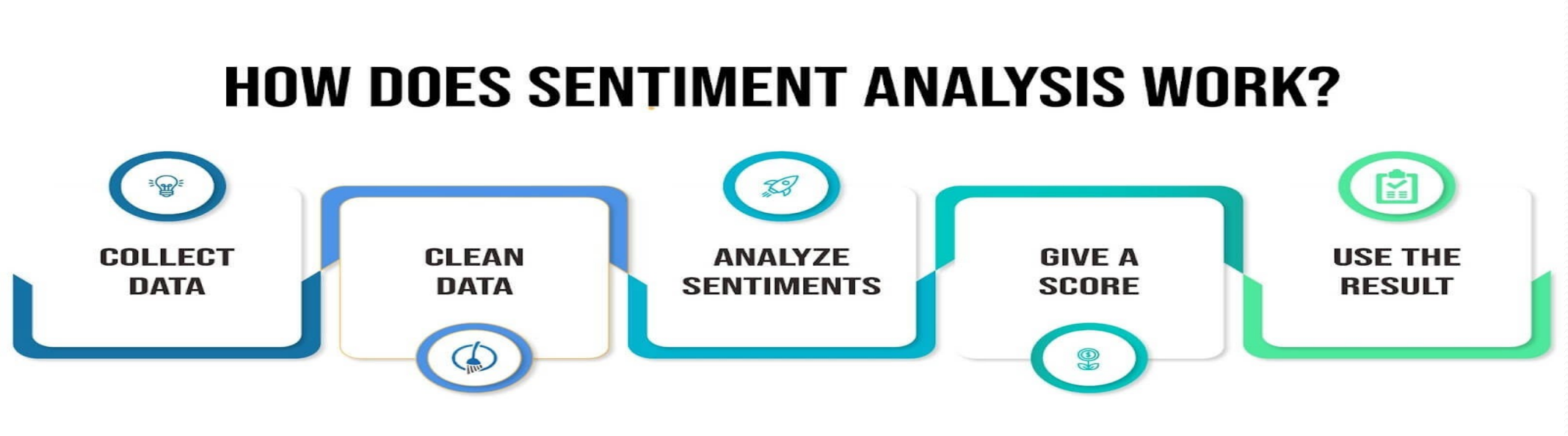
1. GUI Creation: tkinter provides a set of standard GUI components like buttons, labels, text widgets, and canvas, which you can use to build desktop applications with graphical interfaces.

2. Cross-Platform: tkinter is included with Python installations on most platforms.

3. Integration: tkinter widgets can be integrated with other Python libraries and tools.

Overall, tkinter is a versatile library for creating user-friendly interfaces in Python.

Textblob:



- TextBlob is a Python library for processing textual data, providing simple APIs for common natural language processing (NLP) tasks.

- One of its notable features is sentiment analysis, where TextBlob assesses the sentiment polarity (positive, negative, neutral) of a piece of text.

- This is useful for applications like social media monitoring, customer feedback analysis, and sentiment-based recommendation systems.

Overall, TextBlob is favored for its simplicity and ease of integration, making it ideal for prototyping and developing applications that involve textual data processing and analysis.

ReportLab:

- ReportLab is a Python library that allows developers to generate PDF documents programmatically.
- ReportLab provides a comprehensive set of tools for creating dynamic PDF documents in Python. It allows developers to generate PDFs from scratch or modify existing PDFs by adding text, images,charts, and custom vector graphics.
- This makes it suitable for generating professional reports, invoices, certificates, and other types of documents where precise formatting is required.

- Developers can integrate ReportLab with other Python libraries such as Matplotlib for generating charts and graphs directly within PDF documents. It also supports encryption and digital signatures for securing sensitive documents.

Overall, ReportLab is widely used in industries requiring automated PDF generation, such as finance,healthcare, and legal sectors. Its flexibility and extensive documentation make it a robust choice for projects that involve creating complex, customized PDF documents programmatically.

Sumy:

- Sumy is a Python library used for automatic text summarization. It simplifies the task of extracting key information from large blocks of text by offering various algorithms for automatic summarization.

- It provides a straightforward API for summarizing text programmatically, making it useful for applications requiring efficient content extraction and information retrieval.

- The library allows customization of summarization parameters such as the number of sentences or words in the summary, providing flexibility to adapt to different use cases and requirements.

Data Extraction:
The first step in the model development is data extraction, which involves collecting text data from different sources, including URLs and PDF files.

Text Preprocessing:
Once the text is extracted, it needs to be preprocessed to remove any irrelevant information and to prepare it for summarization and analysis.

Summarization:
Summarization is a critical part of the model development. The goal is to condense the text whileretaining the most important information.

Sentiment Analysis:
Sentiment analysis provides insights into the overall tone of the text. The TextBlob library is used for this task. It calculates the polarity of the text, indicating whether the sentiment is positive, negative, or neutral. This analysis helps users understand the emotional context of the summarized content.

Integration with GUI:
The final step in model development is integrating the text extraction, summarization, and sentiment analysis functionalities with the GUI created using tkinter.

Conclusion & Future Enhancements

Conclusion:

In conclusion, the News Summarizer project represents a significant advancement in the realm of information management and news consumption. By harnessing the power of natural language processing and sentiment analysis, coupled with user-friendly interfaces and robust error handling mechanisms, the project aims to streamline the way users access and digest news content.

Future Enhancements:

Multi-language Support: Expansion of language capabilities to include more languages beyond English, catering to a global audience and increasing accessibility for non-English speakers.

References

- Dipanjan Das Andre F.T. Martins (November 21, 2007)
  This survey emphasizes extractive approaches to summarization using statistical methods.

- Archana AB, Sunitha. C describes comparative study on four different approaches to automatic text summarization. T

- Simran kaur1, wg.cdranil chopra2presented approach towards „k means clustering Automated Text Summarization“.

- Yuanzhi Zhu , Zecheng Xie , Lianwen Jin, Xiaoxue Chen,Yaoxiong Huang and Ming Zhang, “SCUT-EPT : New Dataset and Benchmark for Offline Chinese Text Recognition in Examination Paper” IEEE Access , 2018.

- S.E Benita Galaxy , S. Selvin Ebenzer, “Enhancement of segmentation and zoning to improve the accuracy of handwritten character recognition”

- Khamparia, A., Saini, G., Gupta, D., Khanna, A., Tiwari, S., & Albuquerque, V. H. C. D. (2019).

THANK YOU