Text Summarization Using NLP

[1]Mr.Harshal Bharati, [2]Mr.Ankit Kumar, [3]Mr.Shubham Mate, [4]Mr.Premchand Tarange, [5]Prof. R.C Bhaganangre

[1,2,3,4]Students, [5]Professor Department of Computer Engineering, SKN Sinhgad Institute of Technology and Science, Kusgaon(BK), Lonavala, Pune, India.

[5]Rcbhaganagare.sknsits.comp@gmail.com, [1]Harshalbharati.sknsits.comp@gmail.com

[2]Ankitkumar.sknsits.comp@gmail.com, [3]shubhammate.sknsits.comp@gmail.com

[4]premchandtarange.sknsits.comp@gmail.com

Abstract: This Project represents the work associated with Text Summarization. In this paper, we present a framework for summarizing the massive facts. The proposed framework depends on highlight extraction from internet, using each morphological element and semantic data. Presently, in which large facts is available on the internet, it's far maximum important to offer the improved approaches to extract the statistics quickly and most successfully. It may be very hard for human beings to manually extract the precise of a large document of text. There are lots of text substances to be had on the internet. So, there's a hassle of looking for related files from the quantity of documents to be had and absorbing associated statistics from it. In essence to determine out the previous issues, the automated textual content summarization could be very a whole lot essential. Text Summarization is the method of figuring out the most vital and significant information in a input report or set of related input files and compressing all the inputs into a shorter version preserving its normal goals.

Keywords- Extractive summarization, Machine Learning, Natural Language Processing (NLP),Python , Summarization, Text Summarization,

I. INTRODUCTION

In this paper, we present a framework for Text Summarization. The proposed framework depends on summarizing the text from net, utilizing both morphological factors and semantic records. The period of textual content statistics is growing, and people have less time to examine that information. The Internet, media and other facts assets have a large sell off of records and therefore a gadget is required for producing simpler and brief form of information. So, a tool is required for the customers, which would ease out the attempt for them and to study the entire text or remember. Such a systems or tools would be useful and a exceptional time saver for the users. Hectic schedules made it not possible for all of us to study and get admission to the statistics from News information, biographical statistics or from other journals. A reliable and simpler information is wanted to be green. With summaries, People can make efficient choices right away. The motivation right here is to build such a tool which is efficient and creates summaries automatically. Natural Language Processing (NLP) is a area of computerized cogitation in which PCs probe, recognize, and get significance from human language in a radiant and beneficial manner. In addition to typical word processor jobs that deal with messages, such as basic image placement, natural language processing (NLP) considers the several ways that language is constructed: a few words constitute a declaration, a few declarations constitute a sentence, and sentences reveal ideas. NLP expert John Reeling of the software solutions business "Meltwater Group" explained in his article, "How Natural Language Processing helps Uncover Social Media Sentiment." By breaking down the language Because of its significance, NLP frameworks play a lengthy and comprehensive beneficial representation, controlling punctuation, changing shifting the focus of the talk to the message and translating between dialects as necessary. NLP is used to break down the text so that computers can understand human communication. Fair applications such as programmed message outlines, judgment investigations, topic subdivision, named element acceptance, grammatical feature classification, relationship production, stemming, and the list goes on thanks to this human-PC partnership.

II. HISTORY

The practice of reducing a huge volume of text while maintaining its vital information and general meaning is known as text summarizing. For decades, scholars and practitioners have been interested in this topic because of the necessity to efficiently process large volumes of textual data. To meet the issues of text summarization, numerous methodologies and strategies have emerged throughout time.

Early Methodologies (1950s-1990s): Text summarizing attempts date back to the 1950s, when scholars investigated rule-based systems for extracting essential lines from manuscripts. Hans Peter Luhn pioneered the notion of keyword extraction in the 1960s, where keywords are found and utilized to generate document summaries. With early work on text reduction, sentence extraction,

DOI : 10.35291/2454-9150.2024.0188

and sentence weighting, automated summarization gained traction in the 1970s.

Extractive Summarization (from the 1990s to the 2000s): In the 1990s, the extractive summarizing technique, which includes choosing relevant lines or paragraphs from the original material, gained popularity. To identify sentence relevance, researchers began experimenting with statistical and machine learning techniques, such as word frequency and graph algorithms. Through common objectives and benchmark datasets, the Document Understanding Conference (DUC) and the Text Analysis Conference (TAC) performed critical roles in promoting extractive summarization.

Abstractive Summarization (from the 2000s to the present): Abstractive summarizing, in which the summary is formed by rephrasing and synthesizing text, gained popularity in the 2000s as NLP methods advanced. Techniques such as phrase compression and syntactic parsing were employed in early abstractive methods. In the 2010s, the introduction of neural networks and deep learning revolutionized abstractive summarization, allowing for more fluent and human-like summaries. Pre-trained language models like BERT and GPT-2 have enhanced the quality of abstractive summaries even further.

III. LITERATURE SURVEY

Automatic text summarization approaches (A. T. Al-Taani). In this paper ATS (automated text summarization and the approaches of single document and multi- documents text summarizations have been discussed based on requirements extractive summarization is used.[1]

Automatic text summarizer (A. P. Patil, S. Dalmia, S. Abu Ayub Ansari, T. Aul and V. Bhatnagar). The aim in this paper was to design and construct an algorithm that can summarize a document by extracting key text and modifying this extraction using a thesaurus. Mainly to reduce the size, maintain coherence.[2]

Text Summarization: A Review (S. Biswas, R. Rautray, R. Dash and R. Dash). In this paper the objective is to explore text summarization by using various technologies and methodologies in creating a coherent summary which including the key points of the original input document.[3]

An overview of Text Summarization techniques (N. Andhale and L. A. Bewoor). This paper gives an overview survey on both extractive and abstractive approaches.[4]

Text Summarization: An Essential Study (P. Janjanam and C. P. Reddy). This paper focuses on the study of abstractive text summarization approaches and the state of art machine learning models used to summarize single and multi- documents and eventually leading to large document summarization.[5]

Natural Language Processing (NLP) based Text Summarization (I. Awasthi, K. Gupta, Bhupendra Jogi, S. S. Anand and P. K. Soni).In this paper study of extractive and abstractive text summarization method is done .It uses linguistic and statistical characteristics to calculate the implications of sentences. This paper also aims at less repetition and accurate summary.[6]

A method for abstractive text summarization, aiming to generate concise summaries that capture the main ideas of a text by understanding its content and expressing it in new, condensed phrases. It likely discusses techniques such as natural language processing and evaluation metrics like ROUGE.[7]

This paper provides a comprehensive review of various techniques used in abstractive text summarization. It categorizes methods, discusses their strengths and weaknesses, and analyzes recent advancements in the field, contributing to the understanding of summarization techniques and guiding future research directions.[8]

IV. METHODOLOGY

Step in pre-processing: One procedure that comes before translation is pre-processing. The summarizer system's input is a document or group of linked documents. The document has to be moved into a bag of words or phrases. Natural Language Processing (NLP) stages like phrase segmentation, tokenization, stop word removal, and stemming are included in the pre-processing step. Following the completion of pre-processing, each token's word frequency and reverse document frequency values are determined.

Sentence segmentation: The technique of breaking apart a string of written language into its unit or module sentences is known as sentence segmentation. Languages like English and a few others use punctuation, and using symbols like full stops and period characters are a logical estimation.

Tokenization: Tokenization is the technique of breaking down phrases into separate tokens that are adjusted by the spaces and may be utilized for further understanding and improvement. Discrete words, phrases, keywords, IDs, etc. can all be used as tokens. Tokens or words are separated throughout the tokenization process by line breaks, punctuation, and white space. Depending on the demands, the punctuation or white space may or may not get intertwined.

Elimination of Stop words: Stop Words are terms that are commonly used in the language. Eliminating stop words involves getting rid of terms like "the," "to," "are," "is," and so on. Additionally, stop words are eliminated to improve phrase search support.

Model Architecture: The architecture of the text summarization model consists of several components, including the user interface built using Streamlit and the text processing functionalities provided by the txtai library. Streamlit is used to create a user-friendly interface where users can input text and interact with the summarization model. The interface allows users to customize parameters such as summary length and view the generated summaries in real-time. The txtai library is integrated into the model to perform tasks such as text embedding, similarity calculation, and summarization. This integration enables efficient processing of text data and generation of high-quality summaries.

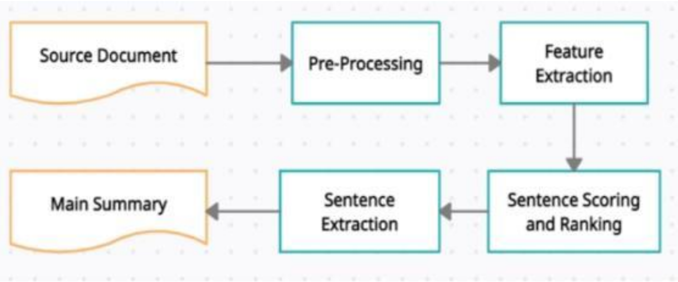


Fig. 4.1 Step wise approach of text summarization

DOI : 10.35291/2454-9150.2024.0188

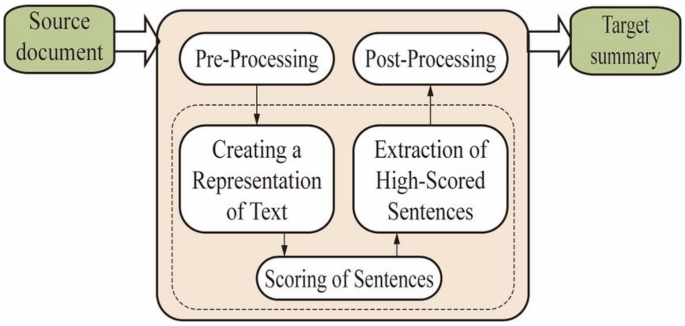


Fig. 4.2 Working

V. Overview of Text Summarization

Text summarization is the process of condensing a large and often complex piece of text, such as an article, document, or news story, into a shorter version while retaining its essential information and meaning. It tries to give readers with a condensed version of the original text, allowing them to comprehend the essential ideas and key points without having to read the full document.

Key Features of Text Summarization:

5.1 Content Reduction: Text Summarization decreases the amount of text by removing unnecessary details, repetitious information, and less significant stuff. This function is especially useful for saving time and improving content consumption.

5.2 Preservation of Meaning*:* Effective text summarizing guarantees that the summary keeps the original text's key meaning and context. This is accomplished by selecting the most significant concepts, ideas, and facts to mention in the summary.

5.3 Machine Learning and Natural Language Processing (NLP): To increase the quality of summaries, several recent text summarizing approaches make use of machine learning and Natural Language Processing (NLP) technology. Extractive summarization has evolved greatly thanks to pre-trained language models such as BERT and GPT-3.

Tools Used

Txtai Library: This library provides various natural language processing (NLP) functionalities, including text summarization. It uses advanced NLP models to understand and condense text into shorter summaries.

Streamlit: This is a Python library used for building web applications. It simplifies the process of creating interactive web interfaces for your Python code. With Streamlit, we can easily design a user-friendly interface for your text summarizer application.

User Interface

Streamlit Interface: Using Streamlit, you design the interface of your application. This can include elements such as text input fields, buttons for uploading files, and areas to display the input text and its summary.

User-Friendly Design: A good user interface design ensures that users can easily navigate and interact with our application. Clear instructions and intuitive controls contribute to a positive user experience.

VI. SYSTEM PROTOTYPE

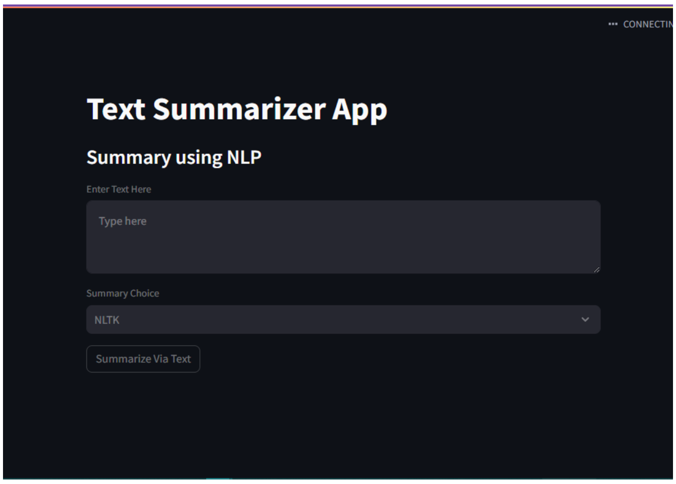
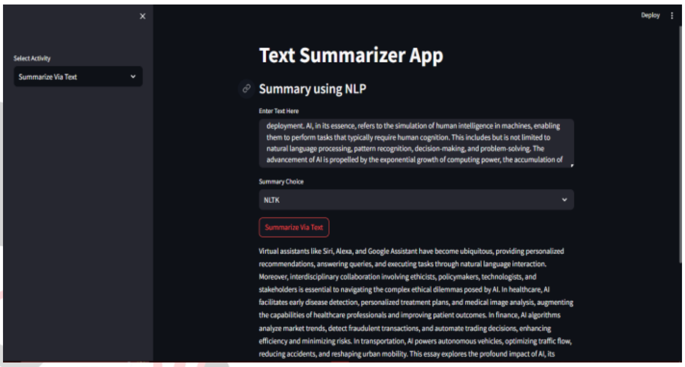


Fig. 5.1 GUI of Application



deployment. AI, in its essence, refers to the simulation of human intelligence in machines, enabling them to perform tasks that typically require human cognition. This includes but is not limited to natural language processing, pattern recognition, decision-making, and problem-solving. The advancement of AI is propelled by the exponential growth of computing power, the accumulation of

Virtual assistants like Siri, Alexa, and Google Assistant have become ubiquitous, providing personalized recommendations, answering queries, and executing tasks through natural language interaction. Moreover, interdisciplinary collaboration involving ethicists, policymakers, technologists, and stakeholders is essential to navigating the complex ethical dilemmas posed by AI. In healthcare, AI facilitates early disease detection, personalized treatment plans, and medical image analysis, augmenting the capabilities of healthcare professionals and improving patient outcomes. In finance, AI algorithms analyze market trends, detect fraudulent transactions, and automate trading decisions, enhancing efficiency and minimizing risks. In transportation, AI powers autonomous vehicles, optimizing traffic flow, reducing accidents, and reshaping urban mobility. This essay explores the profound impact of AI, its

Fig. 5.2 GUI of Application

Functionality of Text Summarization

Text Input: Users can input text into your application through various means, such as typing directly into a text box or uploading a text file.

Summarization Process: Once the text is provided, your application utilizes the text summarization functionality offered by the txtai library. This involves processing the input text using NLP techniques to extract the most important information and generate a condensed summary.

Displaying the Summary: The summarized text is then displayed to the user, typically in a separate section of the web interface. Users can read this summary to quickly understand the main points of the original text.

VII. CONCLUSION

Text summarization is a challenging task in natural language processing (NLP), but it is also a very useful one. It can be used to shorten long documents, making them easier to read and understand. It can also be used to extract key information from documents, which can be helpful for tasks such as question answering and information retrieval. In this project, we have developed a text summarizer using NLP techniques. Our summarizer is based on the extractive approach, which extracts important sentences from the original text to form the summary. We used a variety of NLP techniques to identify important sentences. We evaluated our summarizer on a standard text summarization dataset, and it achieved good results. Our summarizer was able to

DOI : 10.35291/2454-9150.2024.0188

generate summaries that were concise and informative, while still preserving key Information.

REFERENCES

[1] A. T. Al-Taani, "Automatic text summarization approaches," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017, pp. 93-94, doi: 10.1109/ICTUS.2017.8285983.

[2] A. P. Patil, S. Dalmia, S. Abu Ayub Ansari, T. Aul and V. Bhatnagar, "Automatic text summarizer," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 1530-1534, doi:10.1109/ICACCI.2014.6968629.

[3] S. Biswas, R. Rautray, R. Dash and R. Dash, "Text Summarization: A Review," 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), 2018, pp. 231-235, doi: 10.1109/ICDSBA.2018.00048. [4] N. Andhale and L. A. Bewoor, "An overview of Text Summarization techniques," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 2016, pp. 1-7, doi: 10.1109/ICCUBEA.2016.7860024.

[5] P. Janjanam and C. P. Reddy, "Text Summarization: An Essential Study," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1-6, doi:10.1109/ICCIDS.2019.8862030.

[6] I. Awasthi, K. Gupta, Bhupendra Jogi, S. S. Anand and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 1310-1317, doi:10.1109/ICICT50816.2021.9358703.

[7] H. T. Le and T. M. Le, "An approach to abstractive text summarization," 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR), 2013, pp. 371-376, doi:10.1109/SOCPAR.2013.7054161.

[8] P. R. Dedhia, H. P. Pachgade, A. P. Malani, N. Raul and M. Naik, "Study on Abstractive Text Summarization Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020

DOI : 10.35291/2454-9150.2024.0188