e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

EMPOWERING INFORMATION RETRIEVAL: A FRAMEWORK FOR EFFECTIVE DATA SUMMARIZATION USING NLP AND SBERT

Chakradhar Reddy Peddavenkatagari*1

*1Student, Networking and Communications, SRMIST.

ABSTRACT

This project delves into the intricacies of data collection and synthesis, with the primary objective of devising a robust framework for effectively summarizing the extensive knowledge accessible on the Internet. Given the overwhelming nature of the current information landscape, where individuals often struggle to promptly extract pertinent details due to the sheer volume of available data, there is a pressing need for innovative solutions to facilitate efficient information retrieval and assimilation. To address this challenge, the proposed framework harnesses both morphological content and semantic information to sift through and distill relevant data from diverse online sources. Central to this framework is the concept of summary summarization, which entails the identification and condensation of key insights and information from a given dataset into a more concise format, without compromising the original content's essence and purpose. This approach proves invaluable in navigating the complexities of big data, streamlining the information extraction process, and enabling more effective and insightful data analysis.

Keywords- Natural Language Processing (NLP), SBERT, Transformer, Hugging Face, Tokens, Machine Learning, Summarization.

I. INTRODUCTION

This paper presents an extensive framework for Text Summarization, designed to distill relevant information from the vast expanse of data available on the internet by leveraging both morphological and semantic elements. With the exponential growth of text data, individuals are increasingly faced with the daunting task of navigating and comprehending extensive information from various sources such as the internet, media, and other data repositories. This overwhelming volume of data underscores the need for an efficient and user-friendly tool capable of generating concise, easily digestible content to aid in the consumption and understanding of lengthy texts. Such tools are invaluable for individuals with busy schedules, offering an effective solution to save time and enhance productivity.

Summaries play a crucial role in facilitating prompt and informed decision-making by condensing key insights and information from articles, journals, news sources, and biographical content. The objective of this framework is to develop an automatic summarization tool that streamlines the process of extracting essential information, alleviating the manual effort required to sift through extensive content. By offering an efficient and automated solution, this framework aims to improve accessibility and decision-making for users, providing them with the vital information they need without the overwhelming task of navigating through copious amounts of data.

Data analysis involves the selection of key points from articles or documents, an increasingly important task in the face of rising data overload. As the volume of data continues to grow, there is a growing interest in utilizing automated programs to distil relevant information. Two primary methods of article summarization are subtraction and abstraction. Inferential summarization, a form of abstraction, focuses on selecting crucial phrases, sentences, and words from raw text and crafting them into a coherent summary. Abstract summarization, on the other hand, involves creating a concise summary that encapsulates the main points of an article or chapter while maintaining the natural word order aligned with the target sequence or topics.

Natural Language Processing (NLP) stands as a cornerstone in automated cognition, enabling computers to engage with, comprehend, and derive meaning from human language in a sophisticated and practical manner. Through the application of NLP, developers can structure and manipulate data to perform a myriad of tasks, including automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation. NLP empowers machines to interact with human language beyond mere syntax, delving into the semantic nuances and contextual intricacies of communication.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

This capability enables computers to understand not only the words but also the underlying meaning, intent, and sentiment expressed in natural language. For instance, automated summarization involves condensing extensive texts into concise summaries, translation facilitates communication across language barriers, named entity recognition categorizes specific entities mentioned in text, relationship extraction uncovers connections between entities, sentiment analysis gauges the emotional tone conveyed in text, speech recognition enables computers to comprehend spoken language, and topic segmentation organizes content into relevant categories. As technology continues to advance, NLP plays an increasingly pivotal role in enhancing the efficiency and effectiveness of automated systems across various applications and industries, ranging from virtual assistants and chatbots to data analysis and information retrieval.

II. LITERATURE SURVEY

This paper delves into the realm of summarization, specifically focusing on utilizing audio files as the primary input source. These audio files capture human speech, either in real-time or from pre-recorded sources, with users having control over the recording duration through an intuitive GUI equipped with buttons. The recorded speech is saved in WAV format and subsequently converted into a text file, which serves as the input for the text summarization process. The end result is a condensed summarized text file that effectively captures the essence of the initial recording. The project offers users the capability to summarize content sourced from any HTTP link or text data, effectively condensing large volumes of information into concise summaries. The methodology employed involves identifying the highest frequency words in a paragraph, assigning sentence scores based on these frequencies, and selecting the sentence with the highest score to generate the desired summary.

Currently, the system relies on word frequency for its summarization process. To enhance the results of text summarization, the paper integrates techniques utilizing the Gensim library in Natural Language Processing (NLP). These techniques facilitate a deeper and more comprehensive understanding of the overall meaning of the document, providing an effective means to condense and comprehend large amounts of textual information.

The paper explores both extractive and abstractive summarization techniques to cater to diverse applications. While abstractive summarization, which involves generating new sentences, poses challenges in terms of reproducibility and scalability due to its reliance on substantial language production machinery, the straightforward extraction of sentences has demonstrated satisfactory results across various applications. The project successfully achieves its objective by efficiently condensing input textual data into more compact and summarized results.

Text summaries play a pivotal role in various natural language processing tasks, including question and answer systems, text classification, and data retrieval within computer science. The application of text summarization not only enhances the efficiency of information search, leading to improved access times, but also contributes to unbiased algorithms compared to human interpretation. Commercial capture services leveraging text summary systems empower users to efficiently handle a greater volume of texts, thereby enhancing overall performance. The project introduces an entirely data-driven approach to abstractive sentence summarization. The model, which is straightforward and easily trainable end-to-end, is scalable to handle substantial amounts of training data. Leveraging Natural Language Processing (NLP), the system efficiently converts input text into summarized content by translating data stored in files or available on the World Wide Web, utilizing the Beautiful Soup library for web data extraction. Additionally, the project employs the NLTK Rake library to generate keywords from the information and transforms the summarized content into an MP3 file through the gTTS (Google Text-to-Speech) library. The article addresses the challenging task of abstractive document summarization, shedding light on the expansive domain of Text Summarization. Each component of an Automatic Text Summarizer represents a current research focus, offering opportunities for system improvement in terms of capabilities and performance. The future direction of the field involves exploring transformer methods for summarizing multiple documents, enhancing model accuracy with larger datasets from diverse domains, and expanding the evaluation pipeline to include text quality measurement. Extrinsically, evaluating models based on grammar, structure, and referential clarity is crucial for gaining a better understanding and extracting valuable information from the summarized content.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

III. PROBLEM STATEMENT

The integration of Natural Language Processing (NLP) techniques plays a pivotal role in expediting the extraction of valuable information from vast amounts of text available on the web. By leveraging NLP, the process of text summarization becomes more efficient and user-friendly, enabling users to quickly sift through extensive information and obtain the essential insights they require. This not only saves time but also enhances the overall user experience by providing concise and relevant summaries without the need to manually navigate through copious amounts of data.

In this research paper, our primary objective is to explore and implement text-based summarization techniques that are predominantly sentence embeddings-based. Sentence embeddings offer a powerful way to represent textual data in a dense vector space, capturing semantic meanings and relationships between sentences. By utilizing sentence embeddings, we aim to enhance the quality and accuracy of text summarization by focusing on the underlying semantic content and context of the text.

By adopting a sentence embeddings-based approach to text summarization, we anticipate achieving more nuanced and contextually relevant summaries that better reflect the essence of the original content. This approach will enable users to quickly grasp the key points and main ideas presented in the text, facilitating better comprehension and decision-making.

Through this research paper, we aspire to contribute to the advancement of text summarization techniques by emphasizing the utilization of sentence embeddings in NLP. By doing so, we aim to provide users with a more streamlined and effective means of extracting pertinent information from large volumes of text, ultimately enhancing the efficiency and utility of text summarization in various applications and domains.

IV. METHODOLOGY

The extractive summarization process using Sentence-BERT (SBERT) offers a systematic approach to condensing lengthy texts into concise summaries. This method leverages semantic understanding to identify and prioritize essential information, ensuring the extractive summary retains the original text's core message. Below is an expanded explanation of the steps involved in implementing SBERT for text summarization, accompanied by key bullet points for clarity:

Step 1 - Install Necessary Libraries:

Ensure the installation of essential Python libraries like Hugging Face's Transformers and Sentence Transformers.

Use pip, a Python package manager, to install these libraries:

pip install transformers sentence-transformers

Step 2 - Load SBERT Model:

Select and load a suitable pre-trained SBERT model from available options such as 'bert base-nli-mean-tokens' and 'roberta-base-nli-stsb-mean-tokens'.

Step 3 - Tokenization:

Segment the input text into individual sentences using the chosen SBERT model.

Further tokenize the sentences into words if required by the selected model.

Step 4 - Generate Sentence Embeddings:

Utilize the loaded SBERT model to produce embeddings or vector representations for each sentence.

These embeddings encapsulate the semantic essence of the sentences in a high-dimensional vector space.

Step 5 - Similarity Measurement:

Compute pairwise similarity scores between the generated embeddings.

Employ cosine similarity as a prevalent metric to gauge semantic resemblance between sentences.

Step 6 - Sentence Ranking:

Rank the sentences based on their computed similarity scores.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

This ranking mechanism aids in identifying and prioritizing sentences that encapsulate the most critical information.

Step 7 – Summary of Content:

Select the top-ranked sentences to construct the final extractive summary.

Adhere to a predetermined summary length or size to ensure the summary is concise yet informative.

Considerations for Effective Summarization:

Choice of SBERT Model: Opting for a model that aligns with the specific task requirements and the nature of the input text can significantly impact summarization quality.

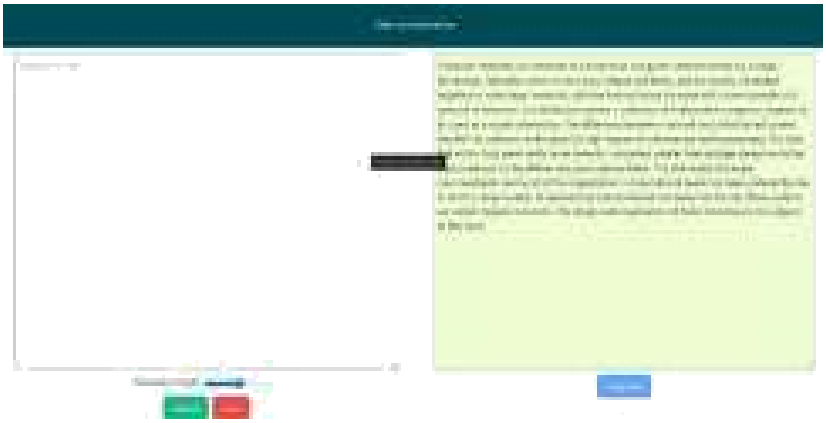
Quality of Pre-trained Embeddings: The efficacy of the summarization process is closely tied to the quality and relevance of the pre-trained embeddings generated by the SBERT model.

Input Text Characteristics: The length, complexity, and topic of the input text can influence the summarization outcome. Adapting the summarization approach based on these factors might be necessary for optimal results.

V. RESULT

Input Data

In these fig – 1 process we are entered the large data on the input section of these framework. Then we have a flexibility to summarization of the data by the help of strength of the button Output Data.



In these fig-2 we seen that large data convert into the small data. It have given the summarization button that the user get above web pages by summarizing the data into specified number of lines based on NLP techniques.

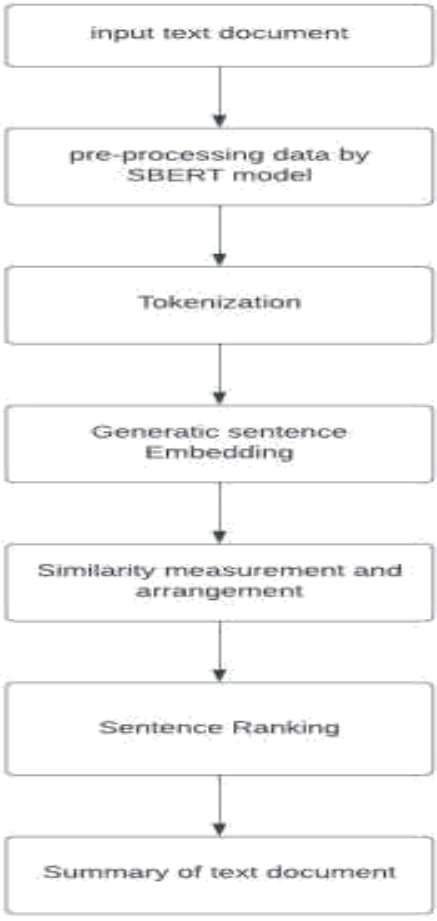
It's essential to consider the specific requirements of your task and adjust the summarization process accordingly. Experimentation with different SBERT models, tokenization strategies, and similarity metrics can help optimize the summarization results for your specific use case. Additionally, fine-tuning the parameters and evaluating the effectiveness of the summarization output are crucial steps in refining the extractive summarization process.

In this research paper, our primary objective is to explore and implement text-based summarization techniques that are predominantly sentence embeddings-based. Sentence embeddings offer a powerful way to represent

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

textual data in a dense vector space, capturing semantic meanings and relationships between sentences. By utilizing sentence embeddings, we aim to enhance the quality and accuracy of text summarization by focusing on the underlying semantic content and context of the text.

By adopting a sentence embeddings-based approach to text summarization, we anticipate achieving more nuanced and contextually relevant summaries that better reflect the essence of the original content. This approach will enable users to quickly grasp the key points and main ideas presented in the text, facilitating better comprehension and decision-making.



VI. CONCLUSION

Incorporating Sentence-BERT (SBERT) into text summarization workflows exemplifies a transformative shift towards harnessing contextually enriched embeddings. This integration not only amplifies the efficiency but also augments the effectiveness of Natural Language Processing (NLP) applications. SBERT's capability to generate embeddings that encapsulate semantic nuances allows for a more nuanced understanding and representation of textual data. As the landscape of NLP continues its rapid evolution, the adoption of SBERT for text summarization emerges as a beacon of innovation.

This approach holds substantial promise in achieving state-of-the-art results, particularly when grappling with the complexities of diverse and dynamic textual content. The versatility of SBERT transcends traditional boundaries, offering a robust framework that researchers and practitioners can leverage. By integrating SBERT, it becomes feasible to elevate the quality and precision of automated text summarization across various domains. Whether it's summarizing news articles, academic papers, or business reports, SBERT's prowess can be instrumental in refining the summarization process, ensuring that the essence of the original content is preserved while eliminating redundancies. In essence, the integration of SBERT into text summarization workflows heralds a new era of NLP-driven solutions, promising more accurate, context-aware, and efficient text summarization capabilities.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

VII. REFERENCES

[1] Pravin Khandare, Sanket Gaikwad, Aditya Kukade, Rohit Panicker, Swaraj Thamke "AUDIO DATA SUMMARIZATION SYSTEM USING NATURAL LANGUAGE PROCESSING" Volume: 06 Issue: 09 | Sep 2019 (IRJET).

[2] M. Monika Rani 2. A. Harika Sweta 3. K Jaswani 4.K Pavan Sidhu 5.Dr.D.N.V.S.L.S Indira "Text Summarization Using NLP" Volume 10 Issue 7 July 2021 IJESI.

[3] Challagundla, Bhavith Chandra. "Efficient CAPTCHA Image Recognition Using Convolutional Neural Networks and Long Short-Term Memory Networks." International Journal of Scientific Research in Engineering and Management (IJSREM), 2024. doi:10.55041/IJSREM29450.

[4] G. Vijay Kumar 1 , Arvind Yadav, B. Vishnupriya, M. Naga Lahari, J. Smriti, D. Samved Reddy "Text Summarizing Using NLP" 2021. [4] Chetana Varagantham, J. Srinija Reddy, Uday Yelleni, Madhumitha Kotha, P. Venkateswara Rao "TEXT SUMMARIZATION USING NLP" Volume 6 Issue 4 August 2022 IJTRET.

[5] AAKASH SRIVASTAVA, KAMAL CHAUHAN, HIMANSHU DAHARWAL, NIKHIL MUKATI, PRANOTI SHRIKANT KAVIMANDAN "Text Summarizer Using NLP (Natural Language Processing)" JUL 2022 IRE Journals Volume 6 Issue.

[6] G. Sreenivasulu, N. Thulasi Chitra, B. Sujatha, and K. Venu Madhav" Text Summarization Using Natural Language Processing" January 2022 ResearchGate.

[7] Balaji N, Deepa Kumari, Bhavatarini N, Megha N, Shikah Rai A, Sunil Kumar P "Text Summarization using NLP Technique" October 2022 ResearchGate.

[8] Challagundla, Bhavith Chandra. "Neural Sequence-to-Sequence Modeling with Attention by Leveraging Deep Learning Architectures for Enhanced Contextual Understanding in Abstractive Text Summarization." International Journal of Machine Learning and Cybernetics (IJMLC), 2024.