JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Data Summerization and Voice Assistant

Prof Mr. Laxman Singh*, Ram Kumar Sharma,Nikhil Saini**,Mrtyunjy Singh****
*Assistance Professor of Computer Science ABES institute of technology, Ghaziabad
**Student, Department of Computer Science ABES Institute of Technology,Ghazibad

***Abstract*— This project focuses on data collection and writing, aiming to develop a framework for efficiently summarizing extensive knowledge available on the Internet. The proposed framework leverages morphological content and semantic information to sift through the vast amount of data online. The current information landscape is overwhelming, making it challenging for individuals to extract pertinent details quickly. The sheer volume of data on the Internet poses difficulties in searching for and assimilating relevant information from diverse sources. The solution lies in the development of an automatic writing system to address these challenges effectively. Summary summarization, a crucial aspect of this framework, involves identifying and condensing the most important and useful information from a given dataset. The goal is to create a concise version while retaining the full purpose of the original data entry. The significance of this approach becomes apparent in the face of the daunting task of making sense of big data, streamlining the process and facilitating efficient extraction of meaningful insights.**

Keywords*— *Natural Language Processing (NLP), SBERT, Transformer, Hugging face, Tokens, Machine Learning, Summarization.

I. INTRODUCTION

This paper introduces a comprehensive framework for Text Summarization, focusing on distilling information from the internet by leveraging both morphological elements and semantic data. Given the escalating volume of text data, individuals often find themselves pressed for time to engage with this extensive information. The internet, media, and various data sources contain vast amounts of data, necessitating a system that can efficiently generate concise and easily digestible content. In response to this challenge, there arises a need for a user-friendly tool that simplifies the process of reading and comprehending lengthy texts. Such tools would prove invaluable, serving as effective time-savers for users with busy schedules. The demands of hectic routines make it challenging for individuals to access and assimilate information from news sources, biographical content, or other journals. To address this, a reliable and user-friendly information system is essential for optimal efficiency. Summaries play a pivotal role, enabling individuals to make informed decisions promptly. The underlying motivation is to develop a tool that not only streamlines the summarization process but also does so automatically, removing the burden of manually sifting through extensive content. By creating a tool that is both efficient and automatic, this framework aims to enhance accessibility and decision-making for users, providing them with the essential information they need without the overwhelming task of navigating through copious amounts of data.

Data analysis involves selecting key points from an article or document, a task increasingly crucial due to the rising problem of data overload. As data volume grows, there is a heightened interest in utilizing programs to distil information. Interpreting content in the era of big data is challenging due to its labor-intensive and time-consuming nature. There are two primary methods for article writing: subtraction and abstraction. Inferential summarization, a form of abstraction, focuses on selecting crucial phrases, sentences, and words extracted from raw text and crafting them into a written form. The significance of the main sentence arises from the analysis and context it provides The summarization process centers on the sentence, allowing for a comprehensive understanding of the data and the extraction of vital information. Abstract summarization involves creating a concise summary that explains the main points of an article or chapter. This process also includes naturally aligning the word order in the document with the target sequence, referred to as topics.

Natural Language Processing (NLP) represents a field of automated cognition in which computers 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 tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation. NLP essentially empowers machines to interact with human language in a way that goes beyond mere syntax, delving into the semantic nuances and contextual intricacies of communication.

This enables computers to not only understand the words but also grasp the underlying meaning, intent, and sentiment expressed in natural language. For example, automated summarization involves condensing large bodies of text into concise summaries, translation facilitates communication across language barriers, named entity recognition identifies and 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. The multifaceted capabilities of NLP make it an integral component in various applications and industries, ranging from virtual assistants and catboats to data analysis and information retrieval. As technology advances, NLP continues to play a crucial role in bridging the gap between human communication and machine understanding, offering endless possibilities for

enhancing the efficiency and effectiveness of automated systems.

LITERATURE SURVEY

This paper focuses on summarization by utilizing audio files as the primary input. These audio files capture human speech either in real-time or from pre-recorded sources. Users have control over the recording duration through a user-friendly GUI with buttons. The recorded speech, saved in wav format, is subsequently converted into a text file. This text file serves as the input for the text summarization process. Ultimately, the project delivers a summarized text file, condensing the content of the initial recording [1].

This project enables users to summarize content from any HTTP link or text data, effectively condensing large volumes of information into concise summaries. The methodology involves identifying the highest frequency words in a paragraph, assigning sentence scores, and selecting the sentence with the highest score to generate the desired summary. Currently, the system relies on word frequency for summarization [2].

To achieve improved results in text summarization, the paper employs techniques utilizing the Genism library in Natural Language Processing (NLP). The incorporation of these techniques facilitates a more comprehensive understanding of the overall meaning of the document, providing an effective way to condense and comprehend large amounts of textual information [3].

Both extractive and abstractive summarization techniques have been explored in response to the diverse applications available. Abstractive summarization, requiring substantial language production machinery, poses challenges in terms of reproducibility and scalability. In contrast, the straightforward extraction of sentences has demonstrated satisfactory results across a wide range of applications. The project successfully fulfills its purpose by efficiently condensing input textual data into more compact and summarized results [4].

Text summaries prove to be valuable 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. Employing a text summary system, commercial capture services empower users to handle a greater volume of texts efficiently, thereby enhancing overall performance [5].

This project proposes an entirely data-driven approach to abstractive sentence summarization. The model, being straightforward and easily trainable end-to-end, is scalable to handle substantial amounts of training data. Employing Natural Language Processing (NLP), the system efficiently converts input text into summarized content. It accomplishes this by translating data stored in files or available on the World Wide Web, utilizing the Beautiful Soup library for the latter. Additionally, the project generates keywords from the information using the NLTK Rake library and transforms the summarized content into an MP3 file through the gTTS library [6].

This article addresses the challenging task of abstractive document summarization, highlighting the expansive domain of Text Summarization. Each component of an Automatic Text Summarizer is a current research focus, offering opportunities for system improvement in terms of capabilities and performance. The future direction 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 better understanding and information extraction [7].

PROBLEM STATEMENT

The use of NLP can reduce the amount of time required to get the useful information and quickly go through the additional information on the web so the text summarization can easily be done and worked through and helps the user in getting the needed information. Through this research paper we wish to use text-based summarization more sentence embeddings based.

METHODOLOGY

Extractive summarization methods involve selecting and combining sentences or phrases from the source text to create a summary. These approaches aim to identify the most informative and representative parts of the text without generating new content. Here are some common extractive summarization techniques:

1) BERT Sum: Fine-tuning BERT, a pre-trained transformer model, for extractive summarization. Sentence embeddings are generated, and sentences are ranked based on their importance scores.

2) Sentence-BERT (SBERT): Using BERT to generate sentence embeddings, SBERT allows for measuring sentence similarity, which can be used to identify important sentences

Sentence-BERT (SBERT) is primarily an extractive method. It focuses on generating fixed-size vector representations (embeddings) for sentences in a way that captures their semantic similarity. The key objective of SBERT is to ensure that similar sentences have similar embeddings in a high-dimensional vector space.

While SBERT is not inherently designed for abstractive summarization, it can be used as a component within a summarization system. For example, SBERT embeddings can be employed to measure the similarity between sentences, aiding in the identification of important sentences during extractive summarization.

Using Sentence-BERT (SBERT) for text summarization involves several steps. Below is a general process outlining how you might execute SBERT for text summarization.

Step 1 - Install Necessary Libraries: Make sure you have the required libraries installed. You may need Python, and you can use libraries like Hugging Face's Transformers, Sentence Transformers, and others. You can install these libraries using a package manager like pip:

```
pip install transformers sentence-transformers
```

Step 2 - Load SBERT Model: Load a pre-trained SBERT model. You can choose from available models such as 'bert-base-nli-mean-tokens', 'roberta-base-nli-stsb-mean-tokens', or others. The model choice may depend on your specific use case and the available pre-trained models.

Step 3 - Tokenization: Tokenize your input text into sentences. Depending on your choice of SBERT model, you might also need to tokenize the sentences into words.

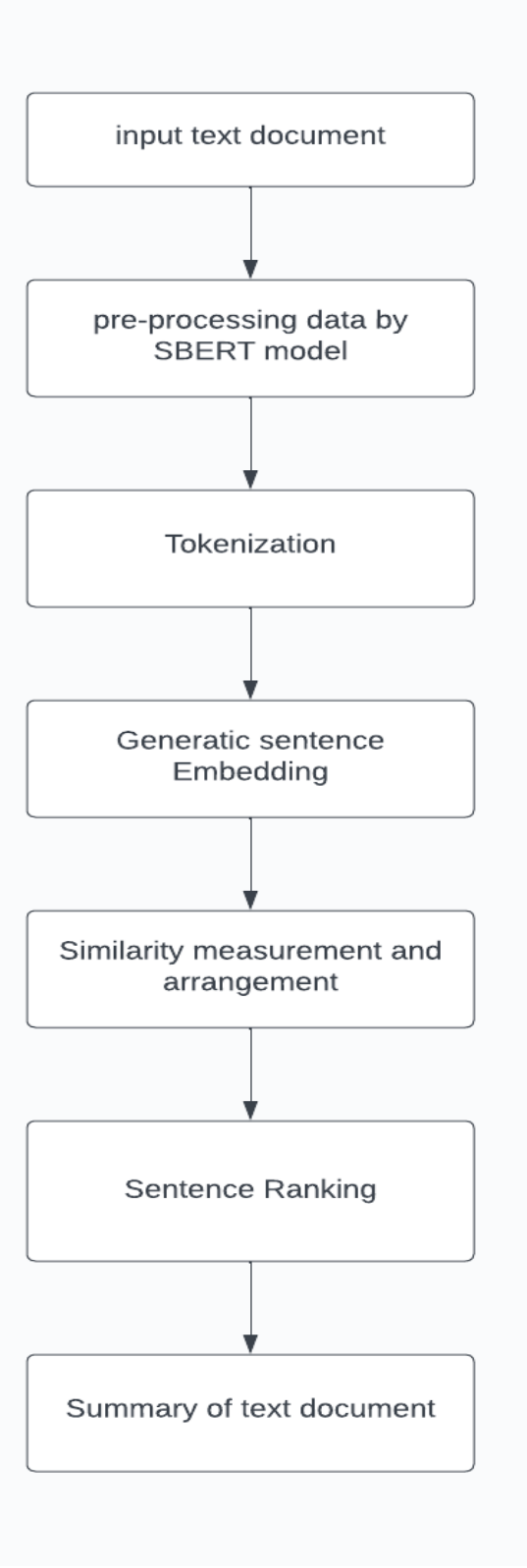
Step 4 - Generate Sentence Embeddings: Use the loaded SBERT model to generate embeddings for each sentence in your input text. The embeddings represent the semantic content of the sentences in a high-dimensional vector space.

Step 5 - Similarity Measurement: Compute pairwise similarity scores between sentences using the generated embeddings. Cosine similarity is a common metric used for this purpose. This step helps identify which sentences are more similar to each other.

Step 6 - Sentence Ranking: Rank the sentences based on their similarity scores. This ranking can be used to prioritize sentences for inclusion in the summary.

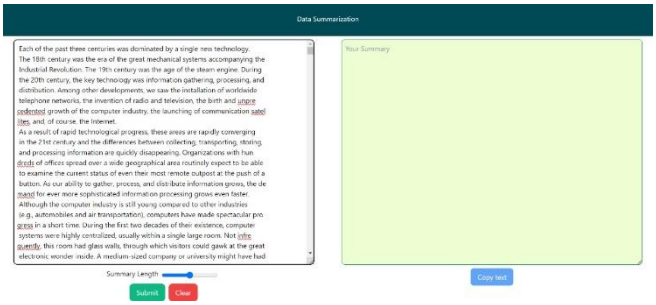
Step7 – Summary of content: Choose the top-ranked sentences based on your desired summary length or a predefined summary size. These sentences form the extractive summary.

It's important to note that the effectiveness of the summarization process may depend on factors such as the choice of SBERT model, the quality of the pre-trained embeddings, and the characteristics of your input text. Additionally, the specific libraries and code structures may vary based on the tools you choose to use.



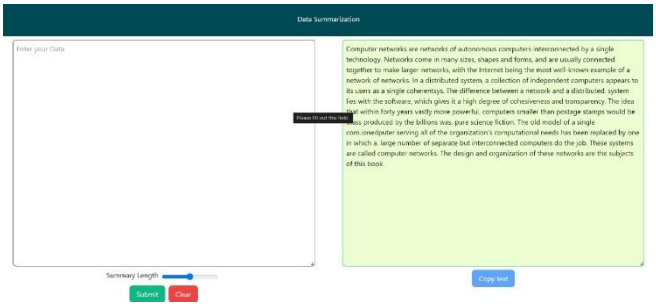
RESULT

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.

CONCLUSION

In the process we are work on the integration of SBERT into text summarization workflows showcases its process in providing contextually rich embeddings, enhancing the efficiency and effectiveness of NLP applications. As the field continues to evolve, leveraging SBERT for text summarization remains a promising avenue for achieving state-of-the-art results and addressing the challenges posed by diverse and dynamic textual content. Researchers and practitioners can harness the power of SBERT to elevate the quality and precision of automated text summarization processes in a variety of domains**.**

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] G. Vijay Kumar 1 , Arvind Yadav, B. Vishnupriya, M. Naga Lahari, J. Smriti, D. Samved Reddy "Text Summarizing Using NLP" 2021.

[4] ChetanaVaragantham,, J.SrinijaReddy, UdayYelleni, 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.