International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-12 Issue-10, September 2023

Data Summerization and Voice Assistant

Laxman Singh, Ram Kumar Sharma, Nikhil Saini, Mrtyunjy Singh

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 quickly extract pertinent details. 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 essential and valuable information from a given dataset. The goal is to create a concise version that retains the original purpose of the 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, a need arises 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.

Manuscript received on 28 August 2023 | Revised Manuscript received on 05 September 2023 | Manuscript Accepted on 15 September 2023 | Manuscript published on 30 September 2023.
*Correspondence Author(s)
Prof Mr. Laxman Singh, Assistant Professor, Department of Computer Science, ABES Institute of Technology, Ghaziabad (Uttar Pradesh), India.
Ram Kumar Sharma*, Student, Department of Computer Science, ABES Institute of Technology, Ghaziabad (Uttar Pradesh), India. Email ID: ram2020cs031@abesit.edu.in
Nikhil Saini, Student, Department of Computer Science, ABES Institute of Technology, Ghaziabad (Uttar Pradesh), India.
Mrtyunjy Singh, Student, Department of Computer Science, ABES Institute of Technology, Ghaziabad (Uttar Pradesh), India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

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 labour-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 summarisation process centres 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

© Copyright: All rights reserved.

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 chatbots 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.

II. LITERATURE SURVEY

This paper focuses on summarisation, utilising 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 summarised text file that condenses 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 summarisation, the paper employs techniques that utilise the Genism library in Natural Language Processing (NLP). The incorporation of these techniques facilitates a more comprehensive understanding of the document's overall meaning, providing an effective way to condense and comprehend large amounts of textual information [3].

Both extractive and abstractive summarisation techniques have been explored in response to the diverse range of applications available. Abstractive summarisation, which requires 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 fulfils its purpose by efficiently condensing input textual data into more compact and summarized results [4].

Text summaries prove 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. Utilising Natural Language Processing (NLP), the system efficiently converts input text into concise, summarised content. It accomplishes this by translating data stored in files or available on the World Wide Web, utilising the Beautiful Soup library for the latter purpose. 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].

A. Problem Statement

The use of NLP can reduce the time required to extract useful information and quickly navigate additional web content, making text summarisation easier to perform and facilitating the user's access to needed information. Through this research paper, we aim to utilise text-based summarisation more sentence embeddings-based.

III. 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 standard 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 essential 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 summarisation, it can be utilised as a component within a summarisation system. For example, SBERT embeddings can be used to measure the similarity between sentences, aiding in the identification of key sentences during extractive summarisation.

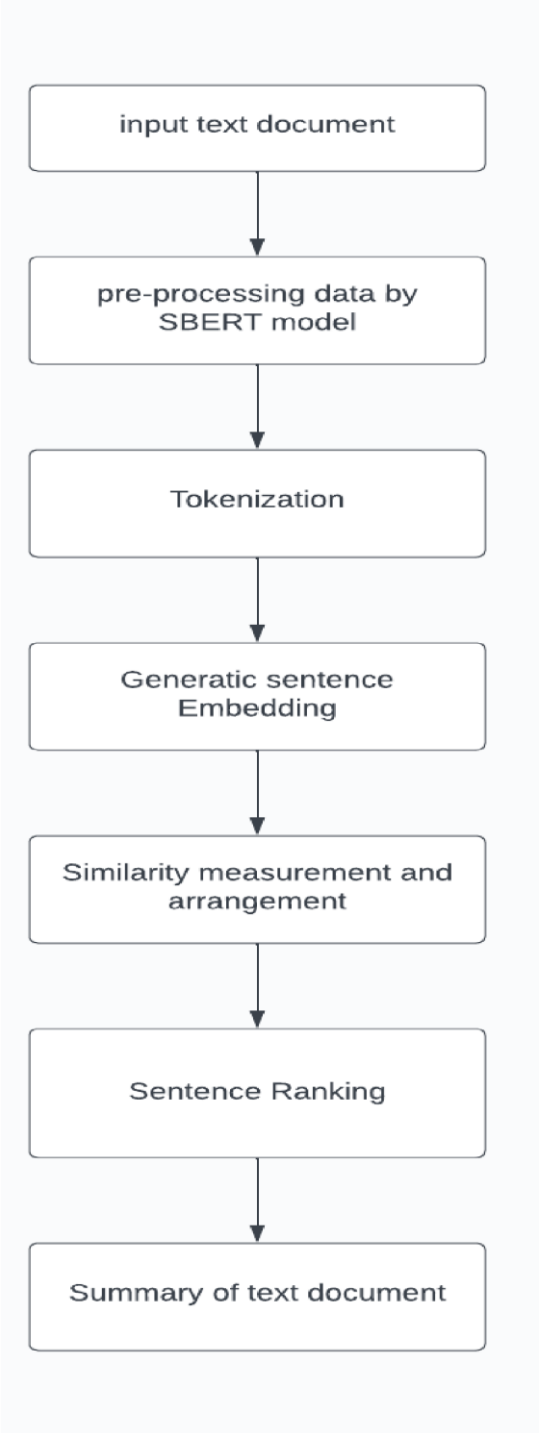
Using Sentence-BERT (SBERT) for text summarization involves several steps. Below is a

Retrieval Number: 100.1/ijitee.C979613030224
DOI: 10.35940/ijitee.C9796.09121023
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

general process outlining how you might execute SBERT for text summarization.

Step 1 - Install Necessary Libraries: Ensure that you have installed the required libraries. You may need Python, and



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 'bertbase-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. you can use libraries like Hugging Face's Transformers and Sentence.

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 standard metric used for this purpose. This step helps identify which sentences are more similar to each other.

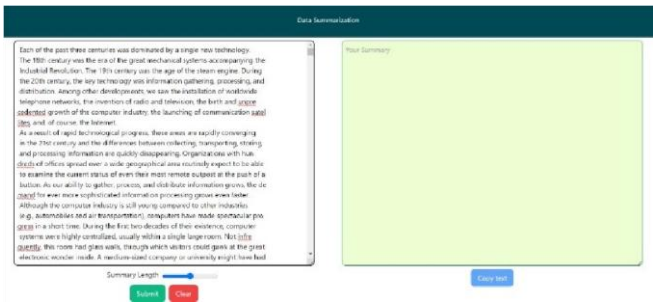
Step 6 - Sentence Ranking: Rank the sentences based on their similarity scores. This ranking can be used to prioritize sentences for inclusion in the summary.

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

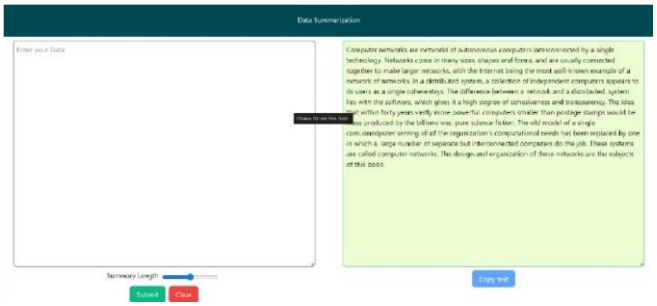
IV. RESULT

A. Input Data



In Fig. 1, we enter the extensive dataset into the input section of the framework. Then we have the flexibility to summarization of the data with the help of the strength of the button.

B. Output Data



In Fig. 2, we see that extensive data is converted into small data. It has a summarisation button that allows users to view the above web pages by summarising the data into a specified number of lines based on NLP techniques.

Retrieval Number: 100.1/ijitee.C979613030224
DOI: 10.35940/ijitee.C9796.09121023
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

V. CONCLUSION

In the process, we are working on integrating SBERT into text summarisation workflows, showcasing its ability to provide contextually rich embeddings that enhance 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 leverage the power of SBERT to enhance the quality and precision of automated text summarisation processes across various domains**.**

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Pravin Khandare, Sanket Gaikwad, Aditya Kukade, Rohit Panicker, Swaraj Thamke "Audio Data Summarization System Using Natural Language Processing" Volume: 06 Issue:
2. 09 | Sep 2019 (Irjet). https://www.irjet.net/archives/V6/i9/IRJET-V6I957.pdf
3. 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. https://www.ijesi.org/papers/Vol(10)i7/Ser-2/B1007020711.pdf
4. G. Vijay Kumar 1, Arvind Yadav, B. Vishnupriya, M. Naga Lahari, J. Smriti, D. Samved Reddy "Text Summarizing Using Nlp" 2021. http://dx.doi.org/10.3233/APC210179
5. Chetana Varagantham, J.Srinijareddy, Uday Yelleni, Madhumitha Kotha, P.Venkateswara Rao "Text Summarization Using Nlp" Volume 6, Issue 4, August 2022. Ijtret. http://dx.doi.org/10.54473/IJTRET.2022.6405
6. 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. https://www.irejournals.com/paper-details/1703633
7. G. Sreenivasulu, N. Thulasi Chitra, B. Sujatha, And K. Venu Madhav" Text Summarization Using Natural Language Processing" January 2022 Researchgate. http://dx.doi.org/10.1007/978-981-16-0739-4_62

AUTHOR'S PROFILE



Asst. Prof Laxman Singh, Assistant Professor of Computer Science and Engineering (CSE) at ABES Institute of Technology (ABESIT), likely possesses strong expertise in CSE, evidenced by his teaching and research contributions. He may hold advanced educational qualifications, fostering a conducive learning environment through practical teaching skills and student mentorship. Additionally, his commitment to the institution and professionalism further enhances his role, potentially supplemented by relevant industry experience.



Ram kumar Sharma is pursuing his B. Tech with CSE from ABES Institute of Technology, Ghaziabad. I completed my schooling at M.V.M Noida. The Computer Science program, with a focus on the Core CSE branch, has provided me with valuable insights and practical knowledge. I am enthusiastic about applying this knowledge to real-world scenarios as I continue to progress in my studies.



Nikhil Saini is pursuing his B. Tech with CSE from ABES Institute of Technology College, Ghaziabad. I completed my schooling at Azad Public School, Bulandshahr. The Computer Science program, with a focus on Core CSE, has provided me with valuable insights and practical knowledge. I am enthusiastic about applying this knowledge to real-world scenarios as I continue to progress in my studies.



Mrtyunjay Singh is pursuing his B. Tech in Computer Science from ABES Institute of Technology, Ghaziabad. I completed my schooling at D.P.S. Rajnagar, Ghaziabad, where the Computer Science program provided me with valuable insights and practical knowledge. I am enthusiastic about applying this knowledge to real-world scenarios as I continue to progress in my studies.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

© Copyright: All rights reserved.