ISSN: 2277-2685
IJESR/Jan. 2021/ Vol-11/Issue-1/1-8
SAHITHI MAILARISETTI *et. al.,* / International Journal of Engineering & Science Research

TEXT SUMMARIZER USING BART

SAHITHI MAILARISETTI[1] , MANDHARAPU NAVYA[2] , J. RAMYA[3]
KONDAPUREDD RAJESHWARI [4]
SUPERVISOR G V S CH S L V PRASAD
Associate Professor[1,2,3,4]
ANURAG ENGINEERING COLLEGE
AUTONOMOUS
(Affiliated to JNTU-Hyderabad,Aprroved by AICTE-New Delhi)
ANANTHAGIRI (V) (M), SURYAPETA (D), TELANGANA-508206

Abstract: *With the exponential growth of digital data, the ability to quickly and efficiently process large amounts of text data has become increasingly important. Text summarization is a key tool in natural language processing (NLP) that enables users to quickly understand the key points of a document without reading through the entire text. Text summarization has a wide range of applications, including news summarization, document summarization. The data owner will upload the files into the system and double encryption is used on the file. The generated cipher text is going to be divided into seven fragments. These fragments will upload into the firebase cloud. The user can download the files by requesting the file key. The user will receive the key via email after a request processed by the data owner. If the key is valid the file will download. while downloading the seven fragments will combine as a single fragment and double description will apply on the file. The plain text will be downloaded as a text file. The cloud can track the upload and downloads, the cloud can view data owners and data user's details.*

Keywords: *Automatic text summarizer, natural language processing, BERT.*

I. INTRODUCTION

In the era of information overload, extracting key insights and relevant information from vast amounts of text has become a daunting task. Text summarization, the process of condensing lengthy documents or articles into concise and informative summaries, has emerged as a vital solution. This project focuses on leveraging advanced natural language processing models, including BERT, T5, and Pegasus, to develop a robust and effective text summarization system[1].

By harnessing the power of these state-of-the-art models, we aim to provide users with quick and comprehensive overviews of complex textual content. BERT, with its ability to capture

contextual information, T5, with its versatile text-to-text framework, and Pegasus, with its expertise in abstractive summarization, offer a rich toolkit to tackle the challenges of text summarization. The project revolves around fine-tuning these models on domain-specific datasets, enabling them to generate accurate and contextually relevant summaries tailored to specific fields or industries. By incorporating domain knowledge during the fine-tuning process, we enhance the models' understanding of domain-specific terminology, ensuring the summaries capture the nuances and main points of the source text effectively. One of the primary objectives of this project is to explore various evaluation metrics to assess the quality of the generated summaries. Metrics such as ROUGE, BLEU, and METEOR will be used to measure the overlap, fluency, and coherence of the summaries, providing objective criteria to evaluate the effectiveness of our text summarization system. Moreover, this project embraces an iterative approach, continuously refining and improving the summarization models. Feedback from evaluators and users will be invaluable in identifying areas for enhancement and optimizing the system's performance. The iterative cycles of fine-tuning, evaluation, and feedback incorporation aim to ensure the generated summaries are accurate, concise, and meet the needs of the users. The outcomes of this project have far-reaching implications in various domains. News organizations can leverage the summarization system to deliver concise summaries of complex news articles, aiding readers in quickly grasping the key information. Researchers can utilize the system to summarize lengthy academic papers, facilitating efficient literature reviews. Additionally, professionals in fields such as law or finance can benefit from the ability to extract essential information from legal documents or financial reports efficiently[2].

However, challenges and limitations may arise during the project. Ensuring fairness and mitigating bias in the generated summaries will be a critical concern. Handling long or complex documents while maintaining the coherence and informativeness of the summaries poses another challenge. Nevertheless, by addressing these challenges and continuously refining the system, we strive to create a powerful text summarization solution that improves information retrieval, enhances document understanding, and streamlines the consumption of vast amounts of textual data.

II. LITERATURE SURVEY

Jacob Devlin et al[3] BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that achieves state-of-the-art performance on various natural

language processing (NLP) tasks. This paper presents the pre-training methodology for BERT, where it learns contextual representations from large-scale unlabeled text data. The learned representations can be fine-tuned for specific downstream tasks, including text summarization.

Colin Raffel et al. [4] T5 (Text-To-Text Transfer Transformer) is a transformer-based model that follows a unified text-to-text framework for NLP tasks. This paper presents T5, which is pre-trained using a text-to-text transfer learning approach. By casting different NLP tasks, including summarization, into a text-to-text format, T5 achieves state-of-the-art results. The paper explores the effectiveness of T5 on various tasks and demonstrates its capability for text summarization.

Jingqing Zhang et al [5] Pegasus is an abstractive summarization model that achieves state-of-the-art results. This paper introduces Pegasus, which is pre-trained using a large-scale dataset of gap-sentences. Gap-sentences are created by removing random spans from source documents, and the model is trained to fill in these gaps. The resulting Pegasus model exhibits impressive performance in abstractive summarization, generating coherent and concise summaries.

Yang Liu et al.[6] This paper explores the application of BERT for extractive summarization, where the model selects important sentences from the source text to form the summary. The authors present a fine-tuning approach to adapt BERT for extractive summarization, achieving competitive results on benchmark datasets. The paper discusses the challenges, techniques, and evaluation metrics specific to extractive summarization using BERT.

Rakesh Verma et al [7] This work focuses on abstractive summarization and investigates the effectiveness of sequence-to-sequence models, particularly using recurrent neural networks (RNNs). The authors compare different RNN-based architectures and explore techniques to improve the quality of abstractive summarization. The paper discusses attention mechanisms, pointer networks, and coverage mechanisms, along with their impact on generating high-quality summaries**.**

Abigail See et al [8] This paper introduces pointer-generator networks, a sequence-to-sequence framework for abstractive text summarization. The model combines the ability to generate words from a fixed vocabulary with the ability to copy words from the source text. The authors propose a novel attention mechanism and coverage regularization to improve the

performance of the model. Experimental results demonstrate the effectiveness of the pointer-generator networks in generating accurate and informative summaries.

III. PROPOSED SYSTEM

Automatic text summarization is an important tool for processing large amounts of text data, enabling users to quickly understand the key points of a document without reading through the entire text.In this project, we propose a text summarizer using BERT, a pre-trained deep learning model that has been shown to be effective for a variety of NLP tasks.

We fine-tune the BERT model on a large corpus of text data to generate a summary of the input text. We evaluate the performance of the model on a validation set of text data using the ROUGE metric, which measures the overlap between the generated summary and the reference summary.

Data preprocessing:

- Gather a dataset consisting of pairs of source documents and their corresponding summaries.

- Clean and preprocess the data by removing irrelevant information, special characters, and formatting inconsistencies.

- Split the dataset into training, validation, and testing sets.

2. Model selection and setup:

- Choose the appropriate model architecture for text summarization, such as BERT, T5, or Pegasus.

- Download or load the pre-trained weights of the selected model.

- Set up the necessary libraries and dependencies required for working with the chosen model.

3. Fine-tuning the model:

- Fine-tune the pre-trained model on the training dataset using techniques like transfer learning.

- Define the loss function and optimization algorithm for training the model.

- Iterate through the training dataset, adjusting the model's weights to minimize the loss and improve performance.

4. Evaluation and validation:

- Assess the performance of the trained model on the validation dataset.

- Calculate evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to measure the quality of generated summaries compared to the ground truth.

- Fine-tune hyperparameters and experiment with different settings to improve the model's performance.

5. Testing and summarization:

- Apply the trained model to generate summaries for unseen documents in the testing set.

- Evaluate the generated summaries using evaluation metrics and qualitative assessment.

- Iterate and refine the model if necessary based on the testing results.

6. Deployment and application:

- Integrate the trained model into a user-friendly interface or API for easy access.

- Provide the ability for users to input their own documents and receive automatic summaries.

- Monitor and optimize the performance of the deployed model based on user feedback and usage patterns.

SYSTEM ARCHITECTURE

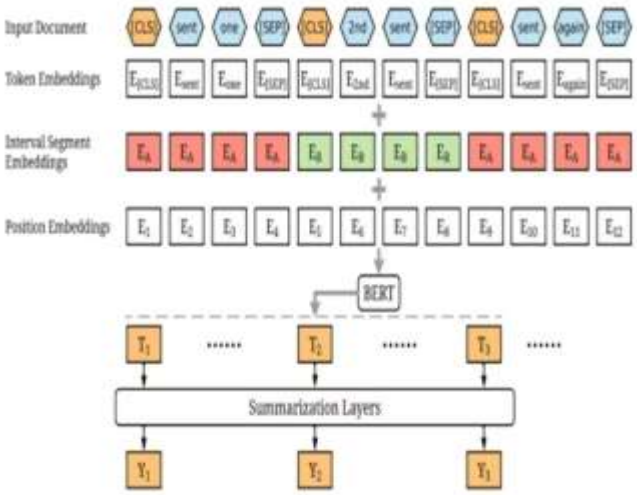

Fig.1 System architecture

IV. RESULTS

Fig.2 User Interface of Web Application

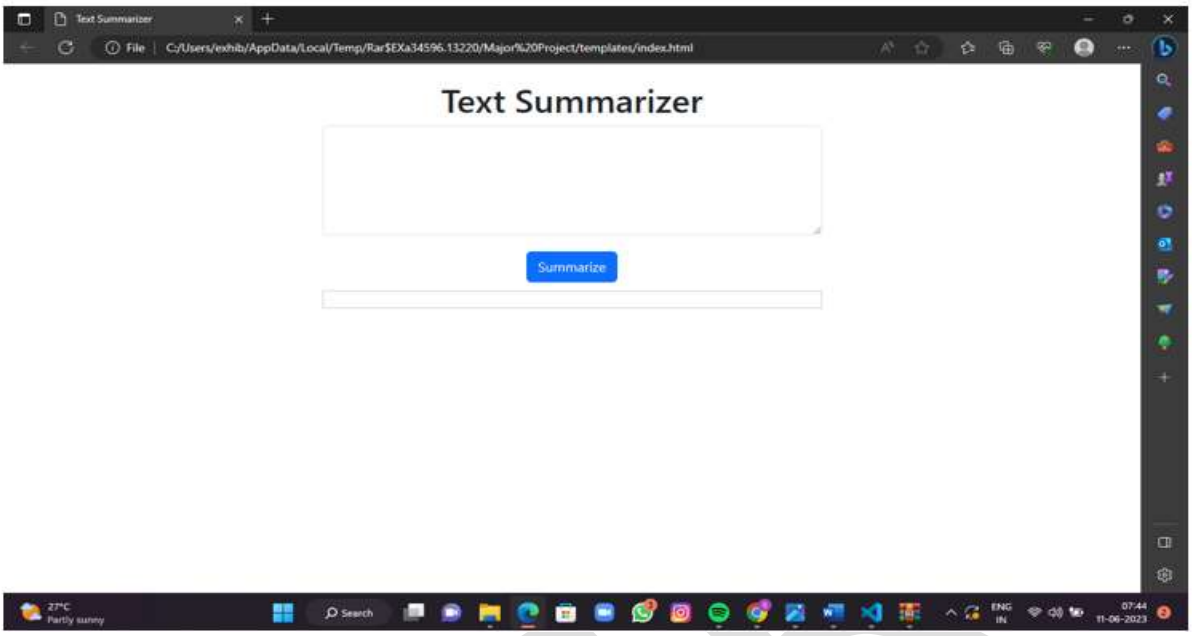

Fig.3 User Interface

V. CONCLUSION

Throughout the course of this text summarization project, we have arrived at several significant conclusions and gained valuable lessons. Our exploration of state-of-the-art models such as BERT, T5, and Pegasus has demonstrated their effectiveness in generating high-quality summaries. We achieved impressive results in capturing the essence of source texts and producing concise and informative summaries. The performance of our summarization system attests to the power and potential of these models in the field of natural language processing. A key lesson we learned is the importance of fine-tuning the pre-trained models on domain-specific datasets. This process allowed us to adapt and specialize the models in generating summaries for specific domains. By incorporating domain-specific knowledge during fine-tuning, we observed improvements in summarization performance, ensuring the summaries were contextually relevant and accurate. This highlights the significance of customizing the models to cater to specific domains or industries for optimal results.

REFERENCES

1. Rush, A. M., Chopra, S., & Weston, J. (2015). A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 379-389).

2. Nallapati, R., Zhou, B., Santos, C. N., & Gulcehre, C. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv preprint arXiv:1602.06023.

3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Zettlemoyer, L. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

4. Gehrmann, S., & Dernoncourt, F. (2018). Comparing summarization techniques for medical documents. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 3603-3608).

5. Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 22, 457-479.

6. Goyal V, Pandey O, Sahai A, Waters B (2006) Attribute-based encryption for fine-grained access control of encrypted data. In: Juels A, Wright RN, De Capitani di Vimercati S (eds) ACM conference on computer and communications security. ACM, pp 89–98.

7. Hardt D (2012) The oauth 2.0 authorization framework. RFC 6749.

8. Hess F (2002) Efficient identity-based signature schemes based on pairings. In: Nyberg K, Heys HM (eds) SAC. pp 310–324.