Enhancing Text Summarization

Devesh Kumar Yadav[1], Balwant Kumar[2], Shachi Mall[3]
[1,2]*Department Artificial Intelligence and Data Science, School of Computer Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*
[3]*Professor, Department Data Science, School of Computer Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

***Abstract*—In the proposed project, an extensive analysis and implementation of different text summarization architectures using deep learning techniques have been provided. In this context, it is mentioned that the proposed system includes an extensive analysis and implementation of different text summarization techniques using deep learning techniques. In this context, it is mentioned that the proposed system includes an extensive analysis and implementation of different text summarization techniques using deep learning techniques, including extractive text summarization techniques and abstractive text summarization techniques using transformer architectures. In this context, it is mentioned that in the proposed system, the BERT encoder model, i.e., bert-base-uncased, is used to select the most important sentences in the text using a classification layer over contextual embeddings. Furthermore, in this context, it is mentioned that a custom sequence-to-sequence model using LSTM architecture is used in the proposed system to implement abstractive text summarization techniques using an encoder mechanism and a decoder mechanism with attention techniques to generate text summaries. Moreover, in this context, it is mentioned that state-of-the-art text summarization techniques using T5 and BART architectures, i.e., transformer architectures, are used in the proposed system to generate text summaries. As per the experimental analysis, it is concluded that the proposed system using state-of-the-art architectures outperforms other text summarization techniques in terms of fluency, coherence, and contextual understanding. Thus, it is concluded that the proposed project includes an extensive analysis and implementation of traditional deep learning architectures using state-of-the-art architectures.**

***Index Terms*—BERT, LSTM, Deep Learning, Text Summarization and Transformer.**

I. INTRODUCTION

A. Background and motivation

The domain of text summarization has evolved as an important area of research over the past few years, with significant challenges associated with quickly understanding the overall idea or gist of the text. Considering the large amount of information readily available today, it is often time-consuming to grasp the overall idea or general concept of the text. The traditional techniques often fail to achieve the purpose, with simplistic ideas not considering the deeper meanings associated with the text. The deeper meanings associated with the text, including the interconnection between the different sentences, play an important role in creating an efficient summarization system. Thus, it is not just about choosing the right words, but it is about understanding the text. With significant advancements in natural language processing and deep learning, it has been possible to achieve significant results, creating efficient summarization tools. These tools not only allow for the identification of significant sentences within the text, but they can also generate new sentences that can effectively summarize the text. This is achieved using models such as bert-base-uncased, T5, and BART, considering their ability to understand the deeper meanings associated with the text. There are still certain issues associated with the efficiency of the models.

B. Problem Statement

This is notwithstanding the fact that the current generation is flooded with information in digital form. The main problem seems to be that the information provided in a summary using basic information processing techniques is not representative of the true

meaning of the source material, and this has been a source of discontent and frustration to users. The fact that users do not have effective information processing techniques at their disposal means that lengthy texts may be overwhelming to read, as they have to sift through superfluous information, and this increases the time spent in the process. There is, therefore, a compelling need to process information using more effective techniques such as bert-base-uncased, T5, and bart.

C. Research Objectives

The major goals and objectives of this assignment are as follows:

1. Development of an automated system for text summarization using deep learning techniques.
2. Development of meaningful text representations using natural language processing techniques in order to improve text understanding.
3. Addressing two major issues:

- Extracting key sentences from a given text using an extractive approach with models like bert-base-uncased.
- Development of an abstract text summarization system using an abstract sequence-to-sequence approach with attention mechanisms.

4. Utilization of pre-trained models like T5 and BART to generate accurate and coherent text summaries in terms of fluency.
5. Development of an accurate system that reduces reading time using text summarization techniques.

D. Contributions of the paper

This paper contributes to the topic in the following ways:

- A method for extractive summarization that is based on natural language processing and utilizes contextual embeddings, such as those created by bert-base-uncased models, to find the most important sentences in the text.
- A method for abstractive summarization that utilizes a sequence-to-sequence model and attention mechanisms to create concise and relevant summaries of the text.
- The use of pre-trained models such as T5 and BART to improve the summarization technique.

II. LITERATURE REVIEW

A. Deep learning for text summarization

Text summarization systems employ deep learning techniques to produce concise and meaningful text in an automated fashion, thus enabling the rapid comprehension of large amounts of text. Conventional text summarization systems employ machine learning techniques, wherein word frequency, sentence positions, and statistical significance are utilized to produce text summaries. These conventional techniques comprise Support Vector Machine (SVM), k-Nearest Neighbor (kNN), decision trees, and random forests. However, the performance of these techniques is largely dependent on the feature selection mechanism. Moreover, these techniques are incapable of extracting semantic meaning. With the advent of deep learning, text summarization systems can be designed using techniques like recurrent neural networks and transformer-based models like bert-base-uncased.

B. Deep learning-based text summarization

Moreover, the advancement of deep learning, especially through the use of neural networks, has greatly enhanced the performance of text summarization systems, especially through their increased level of accuracy and understanding of the text. This can be seen through the use of recurrent neural networks and long short-term memory, which have been shown to have the ability to understand the relationship between sentences, thereby improving the quality of the summary. This advancement can be further seen through the use of the encoder-decoder model, which allows for the training of the model to generate a condensed form of the input text. The use of BERT-base-uncased, T5, and BART models has greatly been adopted for the improvement of text summarization systems. The use of natural language processing, especially through the understanding of contextual meaning, has greatly enhanced the quality of text summarization systems.

C. Limitations of existing methods

The text summarization systems that are already in place might demonstrate excellent results in a controlled environment because they are usually trained and developed using specific datasets that contain well-structured and clean text data. However,

in a real-world scenario, the text data might differ considerably in terms of style, length, and level of noise, resulting in unpredictable and unreliable summarization results. The majority of the text summarization systems are either extractive or abstractive in nature and do not incorporate the effective aspects of both techniques. This might result in the inability of the system to effectively understand the underlying text and the long-range dependencies within it, resulting in incoherent and incomplete summarization results. Moreover, the usability of the system in a real-world scenario is restricted because it usually fails to incorporate essential factors such as scalability and efficiency.

D. Role of text preprocessing and feature normalization

Text preprocessing is an important step in order to enhance the effectiveness of the text summarization system. The techniques that are included in the process of text preprocessing are cleaning and normalizing the text data, tokenization, lowercasing, and removing extra textual symbols. This process is helpful in providing a standardized input to the deep learning models. The implementation of the text normalization process and proper text formatting is helpful in representing the text data in a standardized and meaningful way, thereby enhancing the process of text summarization. Other techniques that are included in the process of preprocessing the text data are truncation and padding, which are helpful in providing a standardized input to the deep learning models, such as bert-base-uncased. Similarly, other models, such as T5 and BART, also benefit from the standardized input provided in the process of text preprocessing.

III. SYSTEM ARCHITECTURE

The structure of the text summarization system consists of different components: user input, preprocessing, deep learning-based summarization models, and output generation. The input data is fed into the system by passing it through the preprocessing module. The input data is normalized, tokenized, and structured in the preprocessing module. This process prepares the data to pass it through the models and ultimately helps in producing an accurate and precise summary. The main function of the text summarization system is carried out by using different deep learning-based summarization models. For extractive summarization, a transformer-based model named bert-base-uncased is used to create context embeddings and to extract key sentences from the input data. For abstractive summarization, a sequence-to-sequence model using the LSTM mechanism and an attention mechanism is employed to create concise summaries by learning the semantic relationships between the input data. The system also uses different transformer-based models, named T5 and BART, to create coherent, precise, and accurate summaries. These models have the ability to extract both local and global contextual information from the input data and thus help in creating precise summarizations.

Here is the image to show system architecture.

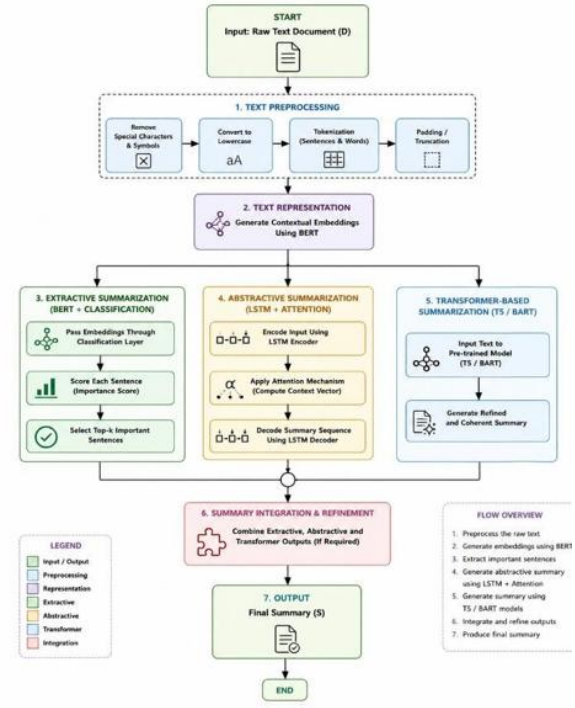

Fig 1: System Architecture

A. Backend and model processing

The backend part is responsible for the core computation of the summarization system. It takes in the provided text from users, processes it into a tokenized format, and then uses this format to pass it through various models that have been set for summarization. In the process of extractive summarization, contextual embeddings are created and passed through a classification layer. In the process of abstractive summarization, encoder-decoder models are used to process the provided text

sequentially. Pre-trained models, such as T5 and BART, are used to create high-quality summaries. The backend part is responsible for controlling the flow of data through various parts of the system, thus allowing for scalability.

B. Data Sources

The system utilizes a CNN/DailyMail dataset for training and evaluation purposes, where the dataset includes a collection of articles and their corresponding summaries written by humans.

The system processes the input text dynamically, meaning it does not rely solely on pre-computed results. The system can generate a summary of the input text in real-time using a trained/pre-trained model.

IV. PROPOSED METHODOLOGY

The proposed text summarization system will follow a structured pipeline where text is processed, meaningful representations are derived, and finally, a summary is generated. The methodology used is a combination of both extractive and abstractive approaches to enhance its performance.

A. Text acquisition and preprocessing

The first input for this system is a text from the user or a dataset. Preprocessing of the text is required as it can vary. The system will clean, tokenize, normalize, and truncate the input text. In addition, it will pad the text as required for uniformity, especially for a deep learning model.

B. Text representation and data preparation

The processed text is converted into a numerical form using tokenization and embedding techniques. The model bert-base-uncased provides contextual embeddings that represent semantic meanings. This allows the model to understand relationships in words and sentences.

C. Extractive summarization

This extractive methodology identifies salient sentences in the input text. A transformer model is employed to create contextual embeddings, and then a classification layer measures the importance of the sentences. The system then selects the most pertinent sentences to build a summary.

D. Abstract summarization

This abstractive method creates innovative summaries using a sequence-to-sequence long short-term memory network with an attention mechanism. Here, the encoder is utilized to encode the input text, and the decoder is utilized to build the summary word by word. The attention mechanism allows the model to focus on important parts of the input.

E. Pretrained model integration

It also utilizes pretrained transformer architectures like T5 and BART to improve performance. These architectures are trained on large datasets and produce better fluency, coherence, and contextual understanding compared to other models.

V. ALGORITHM

Step 1: Text Preprocessing

- Eliminate unnecessary characters from document D
- Lowercase text
- Tokenize sentences and words
- Normalize the text by using padding and truncation

Step 2: Text Encoding

- Create contextualized embeddings from preprocessed text utilizing BERT
- Form sentence vectors

Step 3: Extractive Summarization

- Pass sentence vectors through classification layer
- Assign importance to each sentence
- Retrieve top-k sentences with high importance scores

Step 4: Abstractive Summarization

- Encode input sentences by LSTM Encoder
- Implement attention mechanisms to obtain the relevant information
- Decode sentences to form a summarized text

Step 5: Transformer-based Summarization

- Input the text to the transformer network (T5/BART)
- Produce summary text

Step 6: Output

- Produce summarization output S

VI. MATHEMATICAL MODEL

1. Text Representation using BERT Each sentence $si$is converted into an embedding vector:

$$Ei = \text{BERT}(si)$$

Where:

- Ei= contextual embedding of sentence si

2. Extractive Summarization (Sentence Scoring)
Each sentence is assigned an importance score:

$$Si = \sigma(W \cdot Ei + b)$$

Where:

- Si= importance score of sentence i
- W, b= learnable parameters
- σ= sigmoid activation function Top-k sentences are selected:
  $S = \text{TopK}(Si)$

3. LSTM Encoder-Decoder Model
Encoder:

$$ht = \text{LSTM}(xt, ht{-}1)$$

Decoder:

$$yt = \text{LSTM}(yt{-}1, ct)$$

Where:

- ht= hidden state
- xt= input at time t
- yt= generated output

$$\alpha_{t,i} = \frac{\exp(\text{score}(h_t, h_i))}{\sum_i \exp(\text{score}(h_t, h_i))}$$

$$c_t = \sum_i \alpha_{t,i} \cdot h_i$$

Where:

- $\alpha t,i$= attention weight
- ct= context vector

5. Transformer Attention (T5/BART)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Where:

- Q, K, V= query, key, value matrices
- dk= dimension of keys

6. Loss Function

$$L = -\sum y \log(\hat{y})$$

Where:

- y= actual output
- $\hat{y}$= predicted output

VII. PERFORMANCE ANALYSIS

The evaluation of the proposed text summarization system is carried out in terms of the quality, coherence, and relevance of the generated summary. Various techniques, including extractive, abstractive, and pre-trained transformers, are considered to evaluate their effectiveness. In the extractive summarization technique, using bert-base-uncased, it is observed that the sentences are identified in an efficient manner but may not be fluid due to the direct selection of sentences from the source text. The abstractive summarization technique, using LSTM with attention, produces fluid summaries but may have issues in terms of accuracy due to less training data. Pre-trained models such as T5 and BART provide the best results in terms of fluency and relevance. The proposed system shows that the transformer-based approach outperforms other traditional and custom-built methods in most cases.

VIII. DISCUSSION

A. Interpretation of results
Based on the proposed system, the results obtained show that transformer-based models perform better than traditional and custom deep learning techniques in text summarization tasks. This is because these models are able to learn local and global contextual relationships within the text, resulting in the creation of accurate, meaningful, and coherent summaries. This is unlike other extractive models, which only rely on selecting sentences from the text. Instead, transformer-based models are able to understand the semantic meaning of the text and, in turn, generate accurate summaries that represent the intent of the text. Moreover, T5 and BART models perform better than other models in text summarization tasks due to the generalization ability of the models, which is attributed to the training of the models using large-scale datasets. By comparing all the models, it is clear that although the custom models play an important role in understanding the mechanisms, the transformer-based models are the best solution.

IX. LIMITATIONS

Despite the proven efficacy of the proposed system, certain constraints are still present. The most

prominent constraint is the handling of very long documents, as the majority of the models are based on the transformer architecture, which has input length boundaries that might result in the truncation of the documents and the loss of essential information. Moreover, the requirement of deep learning models, such as the LSTM architecture, is the availability of large amounts of data and computational power to ensure the model performs competitively. The second constraint is that although the models are highly effective, they might sometimes not result in the expected output or contain minor errors. Moreover, the proposed system is not domain-specific, which might result in the system having variable results for the type of text it is intended to process.

X. CONCLUSION

This research develops a comprehensive framework for summarization of texts using a combination of extractive, abstractive, and even pretrained transformers. This framework shows the benefits of using various methods to improve the effectiveness of the final result. The experiment shows that using the pretrained transformers, such as T5 and BART, produces better results in terms of accuracy, coherence, and fluency, and therefore they are best suited to be applied in practice. This research also shows the importance of using deep learning in the improvement of various NLP operations.

XI. FUTURE WORK

- Improvement in processing long documents using advanced techniques related to transformers, like hierarchical attention mechanisms.
- Integration of user-specific preferences and feedback mechanisms to make it more personalized.
- Exploring hybrid models that can successfully integrate both extractive and abstractive methods to leverage their potential.
- Development and extension of this system to include multi-domain and multi-lingual summarization capabilities.

Improvement in terms of computational efficiency to make it more deployable in resource-constrained environments.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171–4186).

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

[4] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 7871–7880).

[5] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (pp. 3104–3112).

[6] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

[7] Bengio, Y., Goodfellow, I., & Courville, A. (2016). Deep Learning. MIT Press.