IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311058|

Text Summarization for Research Papers using Transformers

Vindhya G B, Sushanth H M, Rekha N

Assistant Professor, Department of Computer Science and Engineering, CIT, Tumkur, Karnataka, India

U.G. Student, Department of Computer Engineering, CIT, Tumkur, Karnataka, India

U.G. Student, Department of Computer Engineering, CIT, Tumkur, Karnataka, India

ABSTRACT: A comprehensive evaluation of the body of prior work, which consists of a vast array of publications, is often the first step in scientific study. Researchers would greatly benefit from automatically summarizing scientific papers, which might expedite this research process. The challenges of summarizing scientific publications differ from those of summarizing other types of literature because scientific papers have a particular structure and need for citation phrases. Additionally, useful knowledge that is not immediately apparent in plain language is frequently contained in algorithmic pseudocode, figures, tables, and other components not commonly found in generic text. Using a strong sequential model (seq2seq) and the Transformer model, our Research Paper Assistant offers a thorough examination of the creation and analysis of a complex research paper summary. In order to demonstrate that our work emphasizes the need for abstractive summary strategies in research paper summarization while acknowledging the shortcomings of extractive methods, we are also comparing extractive and abstractive approaches. Extraction techniques may overlook the intricate structure and content of research articles, leading to a loss of coherence and context.

I. INTRODUCTION

The rapid growth in technology has led to a boom in computational linguistics, particularly in the area of NLP Natural Language Processing, which has created new opportunities for creative solutions in scientific research. A particularly interesting area of application is text summarization. With the influx of massive loads of data, reading and summarizing the data in order to extract meaningful information, becomes a taxing task. Thus, automatic summarization techniques have been researched extensively. Summarization is required in a variety of instances, including emails, news stories, official documentation, etc. Many studies are now being conducted in the area of text summarization. It has been around since the very early days of conception of computational linguistics and is still receiving a lot of interest from the research community, as it is far from being considered a solved problem. [2].For text summarization, two popular approaches are employed: Extractive and Abstractive. While abstractive approaches create new material based on the content of the original text, extractive methods identify and extract important phrases or sentences from the text. However, when it comes to summarising academic or scientific publications, these methods haven't shown to be very successful

II. LITEARTURE SURVEY

There are two types of text summary techniques: extractive summarization and abstractive summarization.[17] To comprehend the meaning of the original text, abstractive summarization necessitates creating whole new terms as well as phrases and sentences. The procedure will ultimately approximate human interpretation, despite the fact that this is a more experimental approach. This method uses statistics to choose and pack text from a givendocument.[15]
By using an extractive summarizing technique, you create a condensed version of the original text by choosing the most significant or highly ranked sentences, phrases, and other elements. Sentences' linguistic and statistical propertie shavea significant impact on their significance.[14]

The study includes a comprehensive survey on extractive text summarizing techniques [18]. Machine learning methodology, text summarization with neural networks, the Tf-Idf method (Term Frequency-Inverse Document Frequency), the LSA method (Latent Semantic Analysis), the Cluster Based method, and query-based summarization are the methods used to extract text summarizing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311058|

Several techniques for abstractive text summarization, including Transformer-based models, GANs, attention-based networks, Sequence to Sequence models, and pointer-generated networks, are covered in another study [19].

Researchers used the Extractive technique to summarize scientific literature in another study [2]. They produced a unique dataset with 10,148 publications in computer science. Both neural sentence encoding and conventional summarization aspects were used in their methodology. It's interesting to note that their models surpassed current summarization techniques by combining neural encoding with additional features. These enhanced models demonstrated their efficacy in summarizing scientific papers with remarkable ROUGE-L scores of 0.438 on the CSPubSum Test set and 0.424 on the CSPubSumExt Test set.

[1] Researchers faced a challenging issue when studying a work that focused on summarizing scientific papers: creating concise summaries while preserving crucial details. LSTM with Attention Mechanism, c2c (Copyto-Copy), and fconv (Fully Convolutional) were the three primary techniques they investigated. These approaches did have drawbacks, though, particularly for lengthy works The LSTM approach received low ratings (R-1: 0.375, R-2: 0.173, R-L: 0.329) for failing to capture the key points. Although it still required improvement, the c2c approach showed some improvement (R-1: 0.479, R-2: 0.264, R-L: 0.418). Although it required further refinement, the fconv technique showed promise (R-1: 0.463, R-2: 0.277, R-L: 0.412).

The paper emphasizes the application of the T2SAM model for Abstractive Text Summarization (ATS) [8]. The transformer used in this approach prioritizes semantics over syntax through a self-attention mechanism. Based on ROUGE scores, it outperformed current techniques on both the DUC-2004 and Inshorts News datasets. This concept is highly adaptable since it may also be used to condense news stories, court documents, and product reviews. It is a good substitute for abstractive text summarization because of its simplicity, which stems from a basic Transformer architecture, which makes it more robust and effective than current methods.

III. METHODOLOGY

A. **Information Gathering**

We constructed our own dataset of 853 rows and two columns, "Text" and "Summary," in order to train the model to provide accurate summaries. Each study paper's text is taken out and saved in the "Text" section. Brief and understandable summaries of each research study are included in the "Summary" section. The dataset greatly improved the models' training results by capturing a variety of nuances and specifications included in research papers. We used the freely accessible SciSummnets Corpus [6] as a model to generate this dataset.

B. **Comparative Analysis**

We use our dataset to compare several Extractive and Abstractive summarization models in order to give a thorough review of all the methods discussed and show how our proposed method is superior. A brief description of the models utilized is given below, and Fig. 2 lists the outcomes of those models.

1) Extractive Summarization: As the name implies, [10] this method concentrates on identifying the most significant or recurring, cursory elements within the text. It doesn't go into great into about comprehending the material. We implemented extractive summarization using the following methods: Term Frequency, LSA, and BERT.

a) Bidirectional Encoder Representations from Transformers is referred to as BERT. [11] BERT enables bidirectional pre-training by using a masked language model, in contrast to existing unidirectional language models. In our implementation, we have generated the study paper summary by applying the pre-trained model's effectiveness to our dataset.

b) LSA: To extract the latent semantic structure of the phrases, the Latent Semantic Analysis approach [12] uses statistical computation. The link between terms and sentences in a given text is captured by implementing an input termfrequency matrix.

c) Term Frequency: Using this approach, we assigned scores to the document's sentences according to normalized word frequencies. The top 25% of the highest-ranked sentences and phrases were combined to create the summary. One of the simplest applications of extractive text summarization is this one.

2) Abstractive Summarization: This method creates sentences by comprehending the context of the text rather than just extracting the most pertinent sentences from a source. [16] It effectively gets around a lot of the problems with extractive summarization techniques. Three approaches were examined for abstractive summarization: the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311058|

Transformer-based model, the sequence-to-sequence model without transformer, and the T-5. This was done to highlight even more how our recommended strategy yields superior outcomes.

3) Abstractive Summarization: This method creates sentences by comprehending the context of the text rather than just extracting the most pertinent sentences from a source. [16] It effectively gets around a lot of the problems with extractive summarization techniques. Three approaches were examined for abstractive summarization: the Transformer-based model, the sequence-to-sequence model without transformer, and the T-5. This was done to highlight even more how our recommended strategy yields superior outcomes.

a) T-5: In the realm of natural language processing, the "Text-to-Text Transfer Transformer" concept has proved groundbreaking.[16] The pre-trained model treats each challenge as a text-to-text task by utilizing the encode-decoder architecture and self-attention layers.

b) Sequence-to-Sequence (Seq2Seq) model without Transformer: This model uses an encoder-decoder architecture with several layers of Long Short Term Memory (LSTM). [4] By storing long-term sequences, LSTM, a form of Recurrent Neural Network (RNN), can increase the model's learning and prediction accuracy.

c) Transformer-based approach: In order to improve training results and capture the substantial intricacies of the dataset, we incorporate a [3] Transformer with feedforward network and self-attention layers, taking into account the shortcomings of the previously discussed Abstractive approaches.

C. Design

Regarding the architecture shown in Figure 1, our proposed method is essentially a Transformer-based model that uses an encoder-decoder architecture for sequence-to-sequence communication and integrates feed-forward networks, Layer normalization, and a self-attention mechanism. activities involving abstractive text summarization.

Following data collection, we pre-process the dataset using a variety of methods, such as tokenization, padding, and text sequence truncation, to guarantee that the input and target sequence lengths for the encoder and decoder models are equal. We use batching and shuffling, masking, and model building to perform positional encoding of the input sequence, which aids the Transformer model in understanding the word order in the sequence. bout this approach, which produces the best outcomes.

D. Similarity of Cosines

Finding the connections between the research papers to have a better understanding of the importance of each paper in the individual's research is another application for the Research Paper Mentor. We use a pre-trained BERT model and a tokenizer to create a cosine similarity technique. An overall representation of the input text is obtained by averaging the token embeddings after the tokenizer has encoded the text and obtained embeddings[9] from the BERT model.

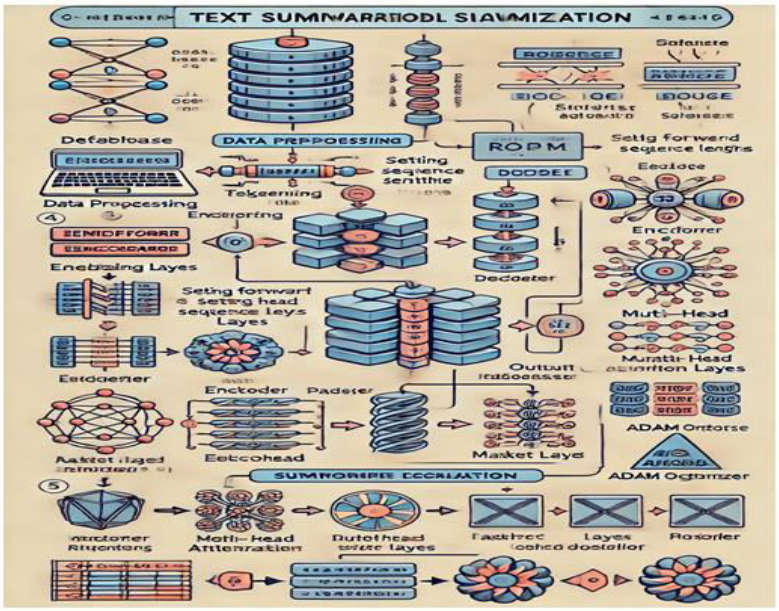

Fig. 1. Architecture of our suggested model

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311058|

IV. RESULTS AND DISCUSSIONS

The rouge (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics commonly used to assess the quality of text summaries and machine translation output. When assessing the caliber of the created summary, they [13] are quite successful. Fig. 2 mentions the findings.

We conducted a comparative study between the Extractive and Abstractive methods of text summarization. We compared three algorithms within Extractive: Term Frequency, BERT, and LSA. T-5, the Seq2Seq model without transformer, and the Seq2Seq model with transformer are all included in Abstractive. It is clear from the given ROUGE scores that the BERT and LSA algorithms perform similarly among the extractive summarization techniques, with corresponding scores of 0.167 and 0.168. However, they perform poorly and generate inaccurate summaries.

This essay methodically explains the issues of prejudice, data security, and morality and value implantation. In this paper the investigations into AI algorithmic ethics highlight the necessity of a multi-faceted approach to address transparency, fairness, privacy, human agency, and governance. Ethical AI development requires collaboration across disciplines to create systems that not only perform efficiently but also align with human values and societal norms.

Fig 2. Comparative results

According to the results of our cosine similarity checker, connected papers had an exceptionally high similarity score (up to 0.93), whilst unrelated papers had noticeably lower ratings. This notable difference in similarity scores shows how well the program works to help users find related publications for their literature reviews.

V. CONCLUSIONS

In order to identify the most efficient method for summarizing academic, scientific, and research publications, the research described in this paper thoroughly evaluates a number of algorithms. The study's findings unequivocally demonstrate that our proposed Abstractive model, which is based on a Transformer-based architecture with an integrated attention layer and an encoder-decoder framework, performs better than any other method examined. Notably, this conclusion is developed and supported by the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, a popular metric used to assess the caliber of text summarization. Additionally, we demonstrate that abstractive models outperform extractive ones when it comes to summarizing scientific papers. The higher ROUGE scores acquired by our suggested model demonstrate its superior performance in capturing and efficiently summarizing the articles from the training dataset. Additionally, our cosine-similarity checker also provides the researcher with an effective method to check whether the research papers studied are relevant to their topic. However, as the results fall

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311058|

short of capturing the important insights and intricate details found in complex academic and research-oriented literature, there is a lot of room for improvement in the summary technique.

REFERENCES

[1] Richard HR Hahnloser, Michael Pfeiffer, and Nikola I. Nikolov. "Data-driven scientific article summarization." (2018) arXiv preprint arXiv:1804.08875.
[2] Sebastian Riedel, Isabelle Augenstein, and Ed Collins. "A method for extractive summarization of scientific papers that is supervised." (2017) arXiv preprint arXiv:1706.03946.
[3] Ashish Vaswani et al. "All you need is attention." Neural Information Processing System Development 30 (2017).
[4] Ibrahim Demir, Xiang, Zhongrun, and Jun Yan. "A rainfall-runoff model using sequence-to-sequence learning based on LSTM." e2019WR025326 in Water Resources Research 56.1 (2020).
[5] Arafat Awajan, Suleiman, and Dima. "Approaches, datasets, evaluation metrics, and challenges for abstractive text summarization based on deep learning." Engineering mathematics problems 2020 (2020): 1–29.
[6] Michihiro Yasunaga et al. "A comprehensive annotated corpus and content-impact models for summarizing scientific papers with citation networks" Proceedings of the AAAI Artificial Intelligence Conference. Volume 33, Issue 1, 2019.
[7] Al-Sabahi K, Nadher M, and Zuping Z (2018) Extractive document summarization (HSSAS) using a hierarchical structured self-attentive model. Access, IEEE 6:24205–24212. 10.1109/ACCESS.2018.2829199 https://doi.org
[8]Solanki, A., and Kumar, S. A transformer model with a self-attention mechanism is used in this abstractive text summarizing technique. 18603–18622 in Neural Comput Applic 35 (2023). 10.1007/s00521-02308687-7 https://doi.org
[9] Derek Miller, "Using BERT to extract instructional text summarization on lectures." (2019) arXiv preprint arXiv:1906.04165.
[10] Gurupreet Singh Lehal, Vishal, and Gupta. "A survey of text summarization extractive techniques." Web intelligence journal of emerging technologies 2.3 (2010): 258-268.
[11] "Bert: Pre-training of deep bidirectional transformers for language understanding" by Devlin, Jacob, and others. (2018) arXiv preprint arXiv:1810.04805.
[12] Darrell Laham, Peter W. Foltz, and Thomas K. Langer. "Latent semantic analysis: an introduction." (1998): 259-284; Discourse processes 25.2-3.
[13] Chin-Yew Lin. "Rouge: An automated summarization evaluation package." The branching out of text summarization (2004).
[14] "A survey on extractive text summarization," by N. Moratanch and S. Chitrakala, International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2017, pp. 1-6, doi: 10.1109/ICCCSP.2017.7944061.
[15] N. Patel and N. Mangaokar, "A Comparative Analysis of Abstractive and Extractive Text Summarization (Output-based Approach)," IEEE International Conference for Innovation in Technology (INOCON), India, 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298416.
[16] "Abstractive Text Summarization using Transfer Learning" by Zolotareva, Ekaterina, Tsegaye Misikir Tashu, and Tomas Horv, et al. ITAT, 2020.
[17]"A survey on extractive text summarization," paper by N. Moratanch and S. Chitrakala, International Conference on Computer, Communication and Signal Processing (ICCCSP), Chennai, India, 2017, pp. 1-6, doi: ICCSP.2017.7944061 10.1109.
[18] "An overview of text summarization techniques," by N. Andhale and L. A. Bewoor, pp. 1–7, 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, doi: 10.1109/ICCUBEA.2016.7860024.
[19]10.1109/ICCCNT54827.2022.9984332 S. Relan and R. Rambola, "A review on abstractive text summarization methods," 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-7.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details