Transformer-based models for generating research highlights from scientific articles

Gowni Bhavishya[1], Siba Sankar Sahu[1]

[1]*Sardar Vallabhbhai National Institute of Technology, Surat, India*

Abstract

Rapid growth of scientific publications makes it difficult for a researcher to identify key contributions. Abstracts usually present a broad summary, while highlights convey the novelty and key findings more precisely. As part of the FIRE 2025 SciHigh shared task, we explore automatic highlight generation, where the goal is to generate concise highlights from scientific articles. In this study, our team SVNIT_CSE explored different transformer-based models for highlight generation. Among the models, BART provide best performance and generate a similar highlight as provided in ground truth. The transformer-based model research highlights generation emphasizes their potential to facilitate faster and more targeted dissemination of scientific knowledge.

Keywords

Research Highlights, Highlight generation, Scientific publications, Transformer-based models

1. Introduction

In recent decades, the rate of scientific publications has increased at an unprecedented rate, with millions of papers being published annually across a variety of disciplines [1]. Rapid growth of scientific publications is a challenging task for researchers, reviewers, and digital libraries to quickly identify the essential contributions of each work. A scientific paper includes an abstract that provides a comprehensive summary of the research work. However, abstracts are often too long, highly descriptive, and stylistically diverse, making them less efficient for rapid screening in extensive literature searches. Hence, a research highlight is essential that comprises brief bullet point statements and summarizes the main contributions of a paper in a clear and structured manner. The research highlights offer immediate information, allowing the reader to assess the relevance information more efficiently. Moreover, it supports digital libraries to improve indexing and retrieval. So far, the researcher has manually generated highlights that have been additional work for authors and editors [2]. Hence, building automated highlight generation systems is essential for research domain that have the capability to produce concise, accurate, and stylistically consistent outputs.

Different studies have been conducted on scientific summarization. In the early days, scientific summarization focused primarily on extractive approaches that extract the most significant sentences from a paper [3, 4]. The extractive approach is simple; but generate poor research highlights. Deep learning models provide better research highlights by using abstractive approaches [5]. The deep learning models have the capability to generate new sentences rather than simply re-create existing ones.

In recent years, transformer models have provided the best performance in different extractive and abstractive summarizations [6, 7]. However, the transformer model is less explored in the research highlight generation. In this study, we explore various transformer-based models to generate research highlights. The explored model provides noticeable performance and outperforms existing models in the generation of research highlights. The research highlights generated by the transformer models are more concise, accurate, and stylistically comparable to those written by the authors. Moreover,

generated highlight balance conciseness with factual accuracy and rarely produce outputs that adhere to the brief, bullet-style format expected of research highlights.

1.1. Task Description

The SciHigh [1] shared task was part of the FIRE [2] (Forum for Information Retrieval Evaluation) evaluation campaign. The goal of the SciHigh shared task is to automatically generate research highlights from scientific articles. In traditional summarization technique, the model generates a summary that is paragraph-long, and it is very difficult to extract relevant information. The model only extracts surface-level key words and provides a summary. The research highlight offers a brief, bullet points and the noteworthy and innovative contributions of a scientific paper. The research highlight focused on creating extremely compressed and information-rich sentences that resemble the highlights of a research paper. This makes the task particularly challenging and requires models to handle domain-specific vocabulary, maintain factual accuracy, and strike a balance between conciseness and informativeness. Highlight generation lies at the intersection of abstractive summarization and scientific information retrieval, with practical applications in enhancing research discoverability.

The remainder of the paper is organized as follows. In Section 2, we present the existing work in text summarization and research highlight generation. Section 3 describes the statistics of the data set. The model framework is presented in Section 4 followed by the experimental result in Section 5. Finally, we conclude with future work in Section 6.

2. Related Work

Text summarization is an important downstream task in the natural language processing domain. Numerous studies have been conducted in high and low-resource languages [8, 9]. Abstractive summarization research has made significant progress with the emergence of sequence-to-sequence models. Nallapati et al. [10] applied recurrent attentional encoder-decoder networks to abstractive summarization. Rush et al. [11] and Chopra et al. [12] implemented sequence-to-sequence models with attention. Different deep neural network models generate novel sentences, but often produce repetitive content or leave out key information. To overcome these limitations, See et al. [5] proposed a pointer-generator network with coverage and achieved a balance between abstraction and the ability to copy technical terms. Later, Anh and Trang [13] improved the model performance by adding Word2Vec and FastText embeddings. These advances marked the emergence of neural abstractive summarization as a viable alternative to extractive methods.

Different researchers also investigated the impact of summarization techniques in the scientific domain [14]. The adaptation of summarization in the scientific domain posed new challenges, including technical vocabulary, long input lengths, and structured discourse. Nallapati et al. [15] introduced a new corpus that comprises an introduction and an abstract section. They explored multiple-time-scale GRUs in arXiv articles. Souza et al. [16] developed a multiview extractive method for scientific texts. Collins et al. [3] presented the CSPubSum dataset [17], which consists of more than 10,000 computer science articles with highlights produced by the authors. Moreover, Cagliero and Quatra [4] refined highlight extraction as a top k sentence selection. Rehman et al. [18] implemented pointer generator networks with GloVe embeddings in CSPubSum [17] for scientific highlight generation. They demonstrated that the proposed methods produced highlights of higher quality than extractive systems. Subsequently, Rehman et al.[19] incorporated named entity recognition into pointer generator models to improve the factual accuracy. They also explored the use of ELMO [20] contextual embeddings to improve semantic representation.

The transformer architecture [21] revolutionizes text summarization by enabling large-scale pre-training. Models such as BERTSUM [22], BART [23], T5 [24], and PEGASUS [25] provide better

---

abstractive summary. Beltagy et al. [26] presented SciBERT, a model pre-trained on academic texts, to address domain-specific language, which enhanced performance of summarization and associated downstream tasks. Most recently, Rehman et al. [19] combined pointer-generator networks with SciBERT embeddings. They showed that abstract-only inputs paired with domain-specific embeddings are sufficient for highlight generation. These transformer-based approaches made the summaries more fluent and logical. They also made the facts more consistent and helped the summaries work better in different domains. As a result, they set a new standard for summarization research and showed that pre-training and fine-tuning methods work for natural language generation.

From the above study, we found that various methods are investigated, such as statistical methods to neural approaches in different domains, such as generic summarization to domain-specific adaptations. However, existing models do not produce accurate, concise, and stylistically appropriate summaries. In this study, we explore different transformer-based models such as BART, T5, LongT5, LED, and PEGASUS that have been widely adopted for general and long-document summarization. These models have not been explored to generate scientific research highlights, which requires highly compressed, contribution-focused output. We evaluated the pre-trained transformer models on the research highlight generation task to establish a baseline of performance and examine their suitability for automated scientific communication.

3. Dataset

In this task, we used the MixSub-SciHigh dataset, a subset of the MixSub corpus [2], which was constructed from research articles published in ScienceDirect. The corpus comprises 19,785 articles in various scientific domains, where each instance is represented as a pair consisting of the abstract of the article and the research highlights written by the authors. The abstracts provide concise summaries of the papers, while the highlights capture their key contributions in short bullet-style statements. The statistic of the MixSub-SciHigh data set is shown in Table 1.

Table 1
Statistics of MixSub-SciHigh dataset

Partition	Number of Instances
Training Set	10,000
Validation Set	1,985
Test Set	1,840
Total	**13,825**

4. Model Framework

We investigated various transformer-based models for scientific highlight generation. All models are evaluated in the MixSub-SciHigh dataset, in which abstracts served as the input and author-written highlights served as the output. Experiments are conducted on Kaggle utilizing an NVIDIA Tesla P100 GPU. In the encoder-decoder models, we experiment with different sequential models and parameters, and the best performance obtained at a given parameter setting is shown in Table 2. We briefly describe the transformer models below.

4.1. BART

Lewis et al. [23] introduced a denoising autoencoder sequence-to-sequence model called BART (Bidirectional Auto-Regressive Transformers). The transformer architecture combines a left-to-right autoregressive decoder, such as GPT, with a bidirectional encoder, such as BERT. Pre-training involves

masking, deletion, and sentence permutation, followed by reconstruction. The architecture of BART is designed in such a way that it offers the best performance in summarization. We experimented with the Facebook/bart-large-cnn[3] variant on the MixSub-SciHigh data set. BART provides the best performance and outperforms all evaluated models.

4.2. T5

Raffel et al. [24] presented the T5 (Text-to-Text Transfer Transformer) model. The model comprises an encoder–decoder architecture and reframes all tasks as text-to-text problems. T5 trained with a 'span corruption' denoising objective. The feature makes it flexible for applications like abstractive summarization and highlight generation. T5 was trained in a large web text data set called the Colossal Clean Crawled Corpus (C4) [24]. We experimented with the `t5-base`[4] in the MixSub-SciHigh dataset. The model provides a fluent and coherent output and offers a longer and less precise summarization.

4.3. LongT5

Guo et al. [27] proposed a variation of the T5 architecture called LongT5. The model has the capability to handle longer input sequences. The conventional T5 model uses the self-attention mechanism, whereas the local-global attention (LoGA) mechanism is incorporated into LongT5. The Local-Global Attention (LoGA) mechanism helps to take a larger input sequence length. The model combines global tokens to collect greater document-level context and local sliding-window attention for short-range dependencies, thus striking a compromise between efficiency and contextual awareness. We experimented with the Google/long-t5-tglobal-base[5] variant on the MixSub-SciHigh dataset. LongT5 takes larger input sequence lengths, but its advantage was limited by the input length.

4.4. LED

Beltagy et al. [28] introduced the longformer encoder-decoder (LED) model, which uses the longformer architecture and is designed for sequence-to-sequence generation. LED combines local attention with a small number of global attention tokens, whereas traditional transformer models fully rely on self-attention. The model can handle token length up to 16,384 tokens, and make the model suitable for long-document summarization. We experimented with the Allenai/led-base-16384[6] variant on the MixSub-SciHigh dataset. The LED architecture is designed to take advantage of long input sequences, but the limited input length in our experiments provides poor effectiveness.

4.5. PEGASUS

Zhang et al. [25] introduced the PEGASUS model. The model was pre-trained using a unique 'gap sentence generation' and specifically designed for abstractive summarization. PEGASUS forces the model to produce sentences from the surrounding context by masking complete sentences during pre-training. PEGASUS-PubMed was pre-trained in biomedical corpora. We experimented with the Google/pegasus-pubmed [7] variant on the MixSub-SciHigh dataset. PubMed performed well, but struggled to generalize outside its biomedical domain. We experimented with different hyperparameters for highlight generation. The best performance achieved by a transformer model in a given hyperparameter is shown in Table 2. The model is trained up to five epochs at a learning rate of 0.00002.

---

[3] https://huggingface.co/facebook/bart-large-cnn
[4] https://huggingface.co/t5-base
[5] https://huggingface.co/google/long-t5-tglobal-base
[6] https://huggingface.co/allenai/led-base-16384
[7] https://huggingface.co/google/pegasus-pubmed

Table 2
Hyperparameters used in different transformer-based models

Model	Batch Size	Max Input Length	Max Output Length
BART	8	384	64
T5	8	512	64
LongT5	8	516	64
LED	4	2048	64
PEGASUS	4	516	64

5. Results and Discussion

In this study, we explore various transformer-based models in highlighting generation from scientific articles. To evaluate the effectiveness of models, we used four widely used evaluation metrics such as ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. The evaluation was carried out on the masked test set (1,840 instances) of the MixSub-SciHigh corpus. Table 3 presents the performance of the transformer models in the generation of highlights. The effectiveness of different transformer models is presented graphically in Figures 1a, 1b, 1c and 1d.

Table 3
Evaluation scores of transformer models on MixSub-SciHigh test set

| Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
BART	34.14	11.97	23.02	26.68
T5	32.74	11.62	22.59	24.26
LongT5	32.64	11.50	22.52	24.15
LED	33.38	11.35	22.50	26.24
PEGASUS	32.30	10.02	21.78	24.35

From the evaluation results, we found that BART provides the best performance and outperforms other models in both ROUGE and METEOR scores. Its pre-training denoising autoencoder, which combines a bidirectional encoder and an autoregressive decoder, contributed to its strong balance of fluency and content coverage. The feature of the BART model is well suited for highlight generation. The effectiveness of each model is varied due to differences in architecture and pre-training objectives. BART achieves comparatively stronger results because its denoising autoencoder objective is well-suited to compressing important information into sharp and short highlights. T5 performs slightly lower, as its general text-to-text formulation is not specifically optimised for abstractive summarisation. The T5 model produced coherent and stylistically fluent outputs. However, the model generates less precise highlights.

The LongT5 architecture has the capability to handle long input sequences due to its extended attention mechanisms. However, these capabilities are not fully utilized in relatively short scientific abstracts, resulting in no notable progress. Moreover, LED is good at handling sparse attention and progressive memory structure that can handle thousands of tokens, but in this task, the short context length does not make the most of its long-range advantages. Both LongT5 and LED were designed to handle much longer input texts, but their advantages could not be fully utilized here due to the shorter abstract lengths. This suggests that long-context architectures may not offer substantial benefits when applied to relatively brief scientific abstracts.
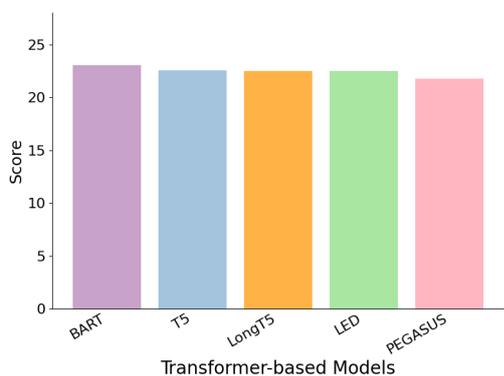
PEGASUS offer competitive performance to other explored models. However, domain mismatch leads to inconsistent and overly descriptive outputs rather than concise highlights. We found that the
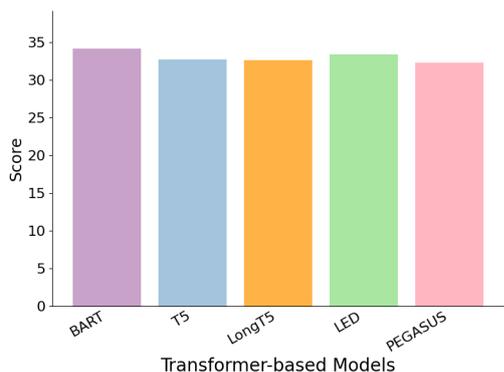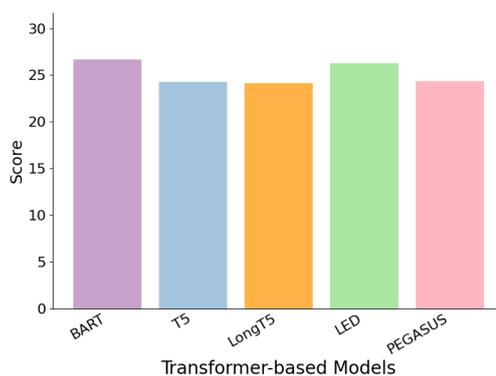
(a) ROUGE-1 scores of different transformer-based models
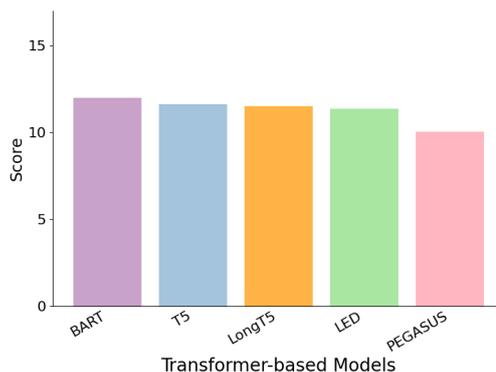
(b) ROUGE-2 scores of different transformer-based models

(c) ROUGE-L scores of different transformer-based models

(d) METEOR scores of different transformer-based models

Figure 1: Performance of transformer-based models based on ROUGE and METEOR metrics

effectiveness of the model is influenced by the pre-training strategy, context handling, and compression ability. It is interesting to note that PEGASUS was trained in biomedical corpora, demonstrating a lack of consistency when applied outside of its target medical domain. Its outputs were often overly descriptive and did not capture concise highlights.

An example of the author-written research highlights and the highlights generated by different transformer-based models is shown in Table 4. The author-written research highlights offer concise and coherent summaries of abstracts. Among the transformer models, BART and T5 produce relatively fluent outputs that preserve key semantic elements. However, BART aligns more closely with the reference highlight. LED and LongT5 attempt to capture technical details, but sometimes generate redundancies or incomplete phrasing. PEGASUS usually focuses on specific details when creating meaningful content. This leads to an incomplete representation of the intended highlight. We found that different transformer architectures are able to balance fluency, content preservation, and conciseness when generating highlights in different ways.

According to the shared task submission policy, we can submit upto two models, and the remaining models we evaluated ourselves. So, we submitted the best performing models BART and T5. Among the models evaluated, BART-large-CNN offers the best performance. The results demonstrate that combining the generative power of large pre-trained transformers with scientific domain enrichment yields highlights that are concise, factually accurate, and stylistically consistent with those written by authors. Table 5 shows the leader board results of the FIRE 2025 SciHigh shared task. Our team, SVNIT_CSE, achieved a ROUGE-L score of 0.2302, securing the third position on the leaderboard. The explored model improves effectiveness and offers competitive performance to other evaluation models in the shared task.

Table 4

Example of author-written (ground truth) research highlights and system-generated highlights

Author-written	Irritability is a transdiagnostic feature of a wide range of psychiatric disorders. Irritability reflects extreme expression of temperamental negative emotionality. Neural substrates of irritability include low midbrain dopamine activity reactivity. Irritability can be expressed as both a mood state and an emotion.
BART	Irritability has transdiagnostic associations with diverse forms of psychopathology. Irritability derives from low tonic dopamine levels and low phasic dopamine reactivity in subcortical neural structures implicated in appetitive responding. Different findings often emerge in neuroimaging studies when irritability is assessed in
T5	We present a model in which irritability derives from low tonic dopamine levels and low phasic DA reactivity in subcortical neural structures implicated in appetitive responding. We propose that more authors use hierarchical Bayesian modeling which capture
LongT5	We describe a model in which irritability derives from low tonic dopamine levels and low phasic DA reactivity in subcortical neural structures implicated in appetitive responding. We propose that more authors use hierarchical Bayesian modeling which capture
LED	I irritability has transdiagnostic associations with diverse forms of psychopathology. In contrast to other emotional states and traits however literature addressing associations between irritability and related temperament and personality constructs is limited. We situate irritability in literatures on child temperament and adult personality and describe a model in which irritability derives
PEGASUS	The of irritability in the literatures on child temperament and adult personality is described. A model suggests that irritability derives from low tonic dopamine levels and low phasic reactivity in subcortical neural structures implicated in appetitive responding. We suggest that more authors use hierarchical modeling which captures functional dependencies between irritability and other

Table 5

Leaderboard results of the FIRE 2025 SciHigh shared task based on ROUGE-L score

Group Name	Run Submission	ROUGE-L	RANK
Text_highlights_gen	run1	0.2345	1
AiNauts	run1	0.2324	2
SVNIT_CSE	**run1**	**0.2302**	**3**
NLPFusion	run2	0.2296	4
The NLP Explorers	run2	0.2294	5
NIT_PATNA_2025	run1	0.2242	6
MUCS	run1	0.2208	7
JU_CSE_PR_KS	run1	0.2206	8
SCaLAR	run1	0.2033	9
Ayanika	run1	0.1791	10
Shilpo	run1	0.1675	11
TJP	run1	0.0673	12

6. Conclusion

The generation of scientific highlight is an important downstream task in the natural language processing domain. In this study, we implemented several transformer-based models, including BART, T5, LongT5, LED, and PEGASUS for highlight generation. Among them, BART outperformed the other models

in the ROUGE and METEOR criteria. The highlight generated by the BART model is quite similar to the user-generated high light. The PEGASUS model has poor effectiveness. Despite these results being encouraging, several challenges remain. The generated highlights are sometimes redundant or not accurate, and current evaluation metrics were not always aligned with human judgments of quality and usefulness. In the future, we can explore the integration of domain-adapted embeddings or retrieval-augmented strategies in high light generation. Additionally, combining automatic metrics with human evaluation will provide a more comprehensive assessment of highlight quality.

Acknowledgments

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Bornmann, R. Haunschild, R. Mutz, Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases, Humanities and Social Sciences Communications 8 (2020) 1–15.

[2] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, P. P. Das, Generation of highlights from research papers using pointer-generator networks and scibert embeddings, IEEE Access 11 (2023) 91358–91374.

[3] E. Collins, I. Augenstein, S. Riedel, A supervised approach to extractive summarisation of scientific papers, in: Conference on Computational Natural Language Learning, 2017, pp. 195–205.

[4] L. Cagliero, M. L. Quatra, Extracting highlights of scientific articles: A supervised summarization approach, Expert Syst. Appl. 160 (2020) 113659.

[5] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, 2017, pp. 1073–1083.

[6] H. P. Luhn, The automatic creation of literature abstracts, IBM Journal of Research and Development 2 (1958) 159–165.

[7] T. Rehman, S. Das, D. K. Sanyal, S. Chattopadhyay, An analysis of abstractive text summarization using pre-trained models, CoRR (2023).

[8] T. Hasan, A. Bhattacharjee, M. S. Islam, K. S. Mubasshir, Y. Li, Y. Kang, M. S. Rahman, R. Shahriyar, Xl-sum: Large-scale multilingual abstractive summarization for 44 languages, in: Findings of the Association for Computational Linguistics: ACL, 2021, pp. 4693–4703.

[9] C. Palen-Michel, C. Lignos, Lr-sum: Summarization for less-resourced languages, in: Findings of the Association for Computational Linguistics: ACL, Association for Computational Linguistics, 2023, pp. 6829–6844.

[10] R. Nallapati, B. Zhou, C. N. dos Santos, Ç. Gülçehre, B. Xiang, Abstractive text summarization using sequence-to-sequence rnns and beyond, in: Proc. 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL, Association for Computational Linguistics, 2016, pp. 280–290.

[11] A. M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, The Association for Computational Linguistics, 2015, pp. 379–389.

[12] S. Chopra, M. Auli, A. M. Rush, Abstractive sentence summarization with attentive recurrent neural networks, in: North American Chapter of the Association for Computational Linguistics, 2016.

[13] D. T. Anh, N. T. T. Trang, Abstractive text summarization using pointer-generator networks with pre-trained word embedding, in: Proceedings of the 10th International Symposium on Information and Communication Technology, 2019, pp. 473–478.

[14] T. Rehman, S. Chattopadhyay, D. K. Sanyal, Abstractive summarization of scientific documents: Models and evaluation techniques, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE, ACM, 2023, pp. 121–124.

[15] R. Nallapati, F. Zhai, B. Zhou, Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3075–3081.

[16] C. M. de Souza, M. R. G. Meireles, R. Vimieiro, A multi-view extractive text summarization approach for long scientific articles, in: International Joint Conference on Neural Networks, IJCNN, IEEE, 2022, pp. 1–8.

[17] E. Collins, I. Augenstein, S. Riedel, A supervised approach to extractive summarisation of scientific papers, in: Conference on Computational Natural Language Learning, 2017, pp. 195–205. Dataset available at GitHub repository as per paper.

[18] S. P. a. T.Rehman, D.K.Sanyal, Automatic generation of research highlights from scientific abstracts, in: Proc. 2nd Workshop Extraction Eval. Knowl. Entities Sci. Documents (EEKE), JCDL, CEUR, Workshop„ 2021, pp. 69–70.

[19] P. M. T. Rehman, D. K. Sanyal, S. Chattopadhyay, Named entity recognition based automatic generation of research highlights, 2022, pp. 163–169.

[20] T. Rehman, D. K. Sanyal, S. Chattopadhyay, Research highlight generation with elmo contextual embeddings, Scalable Comput. Pract. Exp. 24 (2023) 181–190.

[21] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017, pp. 5998–6008.

[22] Y. Liu, M. Lapata, Text summarization with pretrained encoders, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,ACL, 2019, pp. 3728–3738.

[23] N. G. M. G. A. M. O. L. V. S. M. Lewis, Y. Liu, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7871–7880.

[24] A. R. K. L. S. N. M. M. Y. Z. W. L. C. Raffel, N. Shazeer, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, volume 21, 2020, pp. 5485–5551.

[25] J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: Proceedings of the 37th International Conference on Machine Learning, volume 119, 2020, pp. 11328–11339.

[26] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Conference on Empirical Methods in Natural Language Processing, ACL, 2019, pp. 3613–3618.

[27] M. Guo, J. Ainslie, D. C. Uthus, S. Ontañón, J. Ni, Y.-H. Sung, Y. Yang, Longt5: Efficient text-to-text transformer for long sequences, in: Findings of the Association for Computational Linguistics: NAACL, Association for Computational Linguistics, 2022, pp. 724–736.

[28] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, CoRR abs/2004.05150 (2020).