Automatic Scientific Summarization: A Neural Approach to Research Highlight Generation

Ayanika Samanta[1,*,†], Tohida Rehman[1,†]

[1]Techno India University, Kolkata, India.
[2]Jadavpur University, Kolkata, India.

Abstract

With the rapid surge in scientific publications, researchers and indexing platforms increasingly require reliable tools that can condense complex studies into clear and accessible summaries. Research highlights play a particularly important role because they capture the core contributions of a paper in a more focused and digestible form than traditional abstracts. In this work, we introduce a shared task that builds on the previously developed MixSub dataset to automatically generate research highlights from the abstracts of scientific articles. Our objective is to improve the clarity, usefulness, and accuracy of machine-generated highlights so they can better assist academic search and retrieval systems. To explore this task, we fine-tuned transformer-based models, including T5, and evaluated their performance on the shared benchmark. In the SciHigh track at FIRE 2025, we the team *Ayanika* secured the tenth position with a ROUGE-L F1 score of 17.91%.

Keywords

Text Summarization, Natural Language Processing, Pre-trained Language Model, Evaluation Metrics

1. Introduction

The rapid growth of scientific publications across disciplines has made it increasingly difficult for researchers to stay updated with new findings. While abstracts have traditionally served as the primary source of condensed information, they often lack the structural clarity needed for quick scanning and comprehension. In recent years, research highlights typically presented as short bullet points summarizing the core contributions of a study—have emerged as a useful alternative. Their concise and well-structured format makes them particularly valuable for both readers and indexing platforms.

The motivation behind this work is to explore machine learning approaches that can automatically generate such research highlights from scientific abstracts. Unlike general summarization tasks, highlight generation requires identifying key contributions and expressing them in a simplified, point-wise format. To investigate this problem, we use the MixSub dataset, a curated collection of abstract–highlight pairs designed specifically for scientific text generation tasks. This dataset enables systematic benchmarking of different transformer-based summarization methods.

Figure 1 demonstrates a sample abstract–highlight pair from the MixSub dataset, where colors map abstract content to corresponding abstractive author-written highlights.

Using abstracts from MixSub, the shared task focuses on generating structured, compact research highlights that preserve essential information while reducing verbosity.

Additionally, this shared task aims to establish a standardized benchmark that encourages the development of more robust and generalizable highlight-generation models.

The main contributions of this paper are summarized below:

1. We fine-tuned pre-trained transformer models, particularly the T5 text-to-text architecture, to generate structured research highlights and demonstrate their ability to condense lengthy scientific abstracts into concise, meaningful points.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
*Corresponding author.
†These authors contributed equally.
ayanika.samanta1234@gmail.com (A. Samanta); tohidarehman.it@jadavpuruniversity.in (T. Rehman)
0009-0003-2630-2517 (A. Samanta); 0000-0002-3578-1316 (T. Rehman)

2. We evaluate model performance using widely accepted metrics such as ROUGE [1] and METEOR [2], providing a comparative analysis of highlight-generation quality.

Abstract: "Centrosome composed of two centrioles arranged in an orthogonal configuration is an indispensable cellular organelle for mitosis. 130 years after its discovery the structural functional relationship of centrosome is still obscure. Encouraged by the telltale signs of the Mouse and Magnet experiment Paul Schafer pioneered in the research on electromagnetism of centriole with electron microscopy in the late 1960s. Followed by the decades long slow progression of the field with sporadic reports indicating the electromagnetisms of mitosis. Piecing together the evidences we generated a mechanistic model for centrosome function during mitosis in which centrosome functions as an electronic generator . In particular the spinal rotations of centrioles transform the cellular chemical energy into cellular electromagnetic energy . The model is strongly supported by multiple experimental evidences. It offers an elegant explanation for the self organized orthogonal configuration of the two centrioles in a centrosome that is through the dynamic electromagnetic interactions of both centrioles of the centrosome. "

Author-written research highlights:

- ► "We provide a model to describe centrosome function in correlation with its structural organizations."
- ► " We suggested electromagnetic field is the missing link for centrosome function during mitosis."
- ► " We offered physical explanations for the orthogonal self organization structural features of centrosome."
- ► "We provided multiple detailed evidences to support the electromagnetic model we built for centrosome function."

Figure 1: Example of an (abstract, author-written highlight) pair from the MixSub dataset. The same colors in the abstract and highlights indicate which parts of the abstract contributed to each highlight, demonstrating that the highlights are abstractive rather than directly copied. Input and author-written research highlights taken from https://www.sciencedirect.com/science/article/pii/S0303264720301039.

2. Related Work

The rapid expansion of digital information has made automatic text summarization an essential tool for managing and understanding large volumes of content. Early work in summarization relied on statistical and heuristic methods, which selected sentences based on cues such as term frequency, sentence position, or structural markers. As research progressed, extractive systems evolved to incorporate more sophisticated strategies. F To choose the most central sentences, graph-based techniques such as TextRank [3] used algorithms similar to PageRank, in which sentences serve as nodes, and edges denote semantic similarity.

With the advent of deep learning, summarization shifted from simple extraction toward more fluent generation. Abstractive approaches, unlike extractive ones, aim to produce new sentences that capture the underlying meaning of the source text. Transformer-based models such as T5 [4], BART [5], and PEGASUS [6] have further advanced summarization by leveraging self-attention mechanisms and large-scale pretraining, allowing them to better capture semantic relations and long-range dependencies. Sentence embeddings were greatly enhanced by BERT (Bidirectional Encoder Representations from Transformers) [7], which was pre-trained on sizable corpora using masked language modeling. BERTSUM [8] refined BERT to choose important phrases for summaries by adapting its contextual representations for extractive summarizing, despite the fact that BERT was not generative by nature. Abstractive summarization [9], in contrast, aims to construct new phrases that imitate the underlying content, resembling human-written summaries. Templates and language rules were used in early abstractive systems, but these techniques were fragile and domain specific [10]. Abstractive summarization was made possible by the emergence of sequence-to-sequence models with attention [11], which learned mappings between input and output sequences. Nevertheless, long-short-term memory (LSTM) and recurrent neural network (RNN) models frequently generated repeated or insufficient summaries and had trouble handling lengthy dependencies. The transformer design [12] later overcame these drawbacks and served as the basis for contemporary abstractive summarization.

In recent years, attention has increasingly turned toward scientific summarization, which presents unique challenges due to domain-specific vocabulary, technical phrasing, and the need for factual precision. Several studies explore models designed specifically for scientific writing. Rezapour et al. [13] propose a two-stage system that uses structured document representations enriched with scientific graph information, improving both content selection and coherence for long scientific texts.

Rehman et al. [14] used a GRU-based encoder-decoder with Bahdanau attention to build an English text summarizer trained on a news-summary dataset, achieving improved performance for generating concise summaries suitable as headlines. Rehman et al. [15] evaluated pre-trained models such as Pegasus-CNN-DailyMail, T5-base, and BART-large-CNN for summarization across datasets including CNN-DailyMail, SAMSum, and BillSum.

Generating research highlights short bullet points emphasizing key contributions—has emerged as a specialized task within scientific summarization. Early approaches include supervised extractive models [16] and regression-based methods [17], with datasets like CSPubSum, AIPubSumm, and BioPubSumm supporting evaluation. Rehman et al. [18] developed an abstractive pointer-generator model with GloVe, later enhanced with named entity recognition [19], ELMo embeddings [20, 21], and SciBERT with coverage mechanisms, and introduced the multi-domain MixSub dataset [22].

Overall, the literature reflects steady progress across extractive and abstractive techniques, the adoption of transformer architectures, and growing interest in scientific summarization and highlight-generation tasks. Despite this progress, persistent challenges remain, including improving factual consistency, enhancing abstraction, and developing evaluation metrics that more accurately reflect the needs of scientific communication.

3. Methodology

This section presents the transformer-based models considered for fine-tuning on the highlight generation task. Their architectural characteristics and parameter scales are outlined below, and Figure 2 provides an overview of the processing framework.

1. **T5 Family of Models**
   The T5 [4] architecture (Text-to-Text Transfer Transformer) frames every natural language task as a sequence-generation problem . Both the input and the output are treated as text strings, which allows the model to operate within one unified design across multiple applications such as classification, summarization, translation, and question answering. The framework is built on an encoder–decoder transformer, where the encoder produces contextual embeddings and the decoder generates the target sequence autoregressively.
   The T5 model family comes in multiple sizes, from the lightweight T5-small with 60 million parameters to T5-base (220M), T5-large (770M), and the high-capacity T5-3B and T5-11B. Each variant increases the number of encoder and decoder layers, hidden dimensions, and attention heads, providing progressively greater representational power. All versions follow the same architectural design, allowing a trade-off between performance and computational requirements.

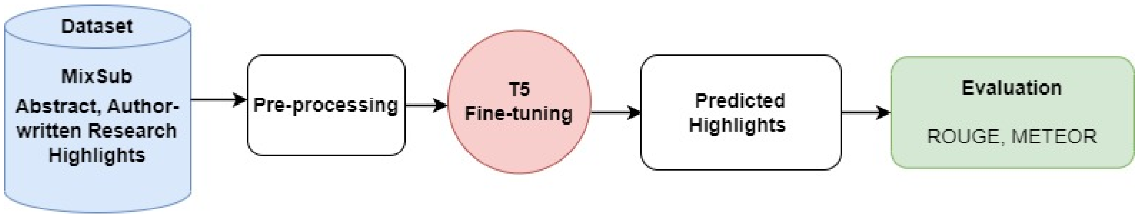

Figure 2: Workflow of the research highlight generation system using fine-tuned transformer models, illustrating the process from abstract input to highlights generation.

4. Experimental Setup

In this section, we discussed the dataset provided for the SciHigh shared task, outline the pre-processing and implementation details, and the evaluation metrics used.

4.1. Dataset

The experiments in this study rely on the MixSub-SciHigh dataset introduced by Rehman et al.[22]. This dataset pairs scientific abstracts with the highlights written by the original authors, making it well suited for training models that aim to generate concise research contributions. In total, it contains 19,785 research articles collected mainly from ScienceDirect and other academic publishers from the year 2020. Each record includes an abstract and its corresponding set of highlights, offering a clear mapping between long-form scientific text and its condensed representation.

Typically, the dataset is divided into training, validation, and test sets using an 80:10:10 split. For the FIRE 2025 SciHigh shared task, a prepared version of the dataset was released, consisting of 10,000 samples for training, 1,985 for validation, and 1,840 for testing. This curated collection serves as the core resource for evaluating systems designed to automatically produce research highlights from abstracts.

4.2. Data Pre-processing

Before model training, several preprocessing steps are applied to ensure that the input text is clean, consistent, and suitable for transformer-based architectures. The process includes the following components:

1. **Data Cleaning:** Removal of extraneous characters, incomplete entries, and formatting inconsistencies to ensure reliable inputs.
2. **Tokenization:** The text is segmented into sentences and tokens using NLTK, allowing the transformer encoder to process the input efficiently.
3. **Normalization:** Standardization steps such as lowercasing, trimming excess whitespace, and reducing redundant punctuation help maintain uniformity across samples.
4. **Abstract–Highlight Alignment:** Each abstract is paired with its corresponding author-provided highlight to create a clear one-to-one mapping for training and evaluation.

4.3. Implementation Details

All model development and experimentation were carried out using the Hugging Face `Transformers` library. The primary model used in this study was `t5-small`[1], chosen for its efficiency and suitability for highlight-style summarization.

The model was trained for **three epochs** on the MixSub-SciHigh dataset. The maximum input length was fixed at **256 tokens**, while the generated highlights were limited to **30 tokens** to maintain concise summaries. These settings ensured consistent training and avoided unnecessary truncation. The model was trained with a learning rate of 2e-5.

[1] https://huggingface.co/t5-small

All experiments were executed on Google Colab using an NVIDIA **T4 GPU**, which was sufficient for full training and evaluation with dynamic padding and standard batching strategies.

4.4. Evaluation Metrics

To assess the quality of the automatically generated research highlights, we employ two widely used metrics in summarization research: ROUGE and METEOR. Both metrics compare system-generated summaries with human-written references, but they capture different aspects of summary quality. ROUGE emphasizes lexical overlap, while METEOR incorporates semantic matching and word-order sensitivity.

4.4.1. ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [1] is a standard evaluation metric for summarization tasks. It measures how much of the reference content is captured by the generated highlight by computing n-gram overlaps. In this work, we report ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 evaluates unigram overlap, ROUGE-2 captures bigram overlap, and ROUGE-L measures the longest common subsequence between the two summaries.

Let O denote the number of overlapping n-grams, N_{gen} the total number of n-grams in the generated highlight, and N_{ref} the total number in the reference highlight. Precision and recall are computed as shown in Equation 1, while the F1-score can be calculated as per the Equation 2.

$$\text{Precision} = \frac{O}{N_{\text{gen}}}; \quad \text{Recall} = \frac{O}{N_{\text{ref}}} \tag{1}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

ROUGE focuses on lexical similarity, providing insight into how well the generated highlight captures the key terms and phrases of the reference summary.

4.4.2. METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [2] offers a complementary perspective by incorporating semantic matching and ordering constraints. Unlike ROUGE, which relies strictly on n-gram overlap, METEOR aligns unigrams using exact matches, stemming, and synonyms, enabling a more meaning-oriented evaluation. It also applies a fragmentation penalty to account for disordered or scattered matches.

Let P represent unigram precision, R represent unigram recall, and let matched unigrams be grouped into ordered *chunks*. The F_{mean} score, fragmentation penalty, and final METEOR score are computed using Equations 3, 4, and 5.

$$F_{\text{mean}} = \frac{10 \times P \times R}{R + (9 \times P)} \tag{3}$$

$$\text{Penalty} = 0.5 \times \left(\frac{\#\text{chunks}}{\#\text{matched\_unigrams}} \right)^3 \tag{4}$$

$$\text{METEOR} = F_{\text{mean}} \times (1 - \text{Penalty}) \tag{5}$$

METEOR provides a more semantically sensitive evaluation by rewarding synonym matches and penalizing disordered alignments, making it well suited for assessing highlight-generation quality.

5. Results

Table 1 presents the F1-score performance of the fine-tuned T5-small model on the highlight generation task. The model achieved a ROUGE-1 F1-score of 26.23%, indicating a moderate level of unigram overlap with the reference highlights. The ROUGE-2 F1-score of 7.91% reflects the expected lower bigram match typically observed in highly abstractive summarization settings. The ROUGE-L F1-score of 17.91% shows that the model was able to capture portions of the longest common subsequence, suggesting reasonable structural alignment with the reference summaries. The METEOR F1-score of 18.71% further supports this observation by indicating a fair degree of semantic correspondence through synonym matching and ordering.

Overall, the evaluation demonstrates that while the T5-small model effectively captures essential terms and preserves parts of the underlying structure, its ability to reproduce multi-word expressions and longer phrase-level dependencies remains limited—an expected challenge in scientific highlight generation. According to the official leaderboard, Team *Ayanika* secured the **10th rank** in the SciHigh track at FIRE 2025 using this model configuration.

Model Name	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
T5-small	26.23	07.91	17.91	18.71

Table 1
F1-score performance of the fine-tuned T5-small model across ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. All scores are reported in %.

As shown in Table 2, our team *Ayanika* achieved a ROUGE-L F1 score of 17.91%, placing them at the 10th position among all participating teams.

Table 2
ROUGE-L F1 performance (percentage values) of the top runs submitted by each team in the SciHigh track at FIRE 2025. Only the best-performing run for every team is reported. All scores are reported in %.

Group Name	Run Submission (Best Run)	ROUGE-L F1	Rank
Text_highlights_gen	run1	23.45	1
AiNauts	run1	23.24	2
SVNIT_CSE	run1	23.02	3
NLPFusion	run2	22.96	4
The NLP Explorers	run2	22.94	5
NIT_PATNA_2025	run1	22.42	6
MUCS	run1	22.08	7
JU_CSE_PR_KS	run1	22.06	8
SCaLAR	run1	20.33	9
Ayanika	**run1**	**17.91**	**10**

5.1. Case Study

To better understand the quality of the generated highlights, we present a case study in Figure 3. This shows an example in which the author-written highlight is compared with the highlight generated by the T5-small model. The case study reveals that the T5-small model correctly captures several core elements from the abstract. Specifically, it identifies key ideas such as "dropout is commonly used to reduce overfitting", "fixed drop probability", and "performance degradation when dropout is applied extensively".

However, the generated highlight remains largely extractive and lacks the concise, contribution-focused expressions found in the author-written highlights. Crucial ideas such as "surrogate dropout, per-neuron drop rate," and "superior regularization performance across datasets" are not captured by the model.

Overall, the comparison indicates that while the T5-small model can identify important surface-level information, it struggles to produce abstracted, contribution-oriented highlights. This highlights the need for improved fine-tuning strategies that strengthen abstraction, compression, and emphasis on novel contributions.

Abstract: "Dropout is commonly used in deep neural networks to alleviate the problem of overfitting. Conventionally the neurons in a layer indiscriminately share a fixed drop probability which results in difficulty in determining the appropriate value for different tasks. Moreover this static strategy will also incur serious degradation on performance when the conventional dropout is extensively applied to both shallow and deep layers. A question is whether selectively dropping the neurons would realize a better regularization effect. This paper proposes a simple and effective surrogate dropout method whereby neurons are dropped according to their importance. The proposed method has two main stages. The first stage trains a surrogate module that can be jointly optimized along with the neural network to evaluate the importance of each neuron. In the second stage the output of the surrogate module is regarded as a guidance signal for dropping certain neurons approximating the optimal per neuron drop rate when the network converges. Various convolutional neural network architectures and multiple datasets including CIFAR 10 CIFAR 100 SVHN Tiny ImageNet and two medical image datasets are used to evaluate the surrogate dropout method. The experimental results demonstrate that the proposed method achieves a better regularization effect than the baseline methods.
Author-written highlight: A simple and effective regularization method called surrogate dropout is proposed which regards the surrogate module as a proxy for approximating the optimal drop rate of each neuron. Compared with conventional dropout the surrogate dropout method has fewer restrictions. Both the shallow and deep layers in CNNs can benefit from the usage of surrogate dropout. The superior regularization effect of surrogate dropout has been empirically verified using multiple datasets and networks with various depths.
Fine-tuned T5-small model generated highlight: Dropout is commonly used in deep neural networks to alleviate the problem of overfitting. Conventionally the neurons in a layer indiscriminately share a fixed drop probability which results in difficulty in determining the appropriate value for different tasks. Moreover this static strategy will also incur serious degradation on performance when the conventional dropout is extensively applied.

Figure 3: Case study comparing author-written highlights and T5-small generated highlights. Input and author-written research highlights taken from https://www.sciencedirect.com/science/article/pii/S0950705120304986.

6. Conclusions and Future Work

Our team *Ayanika* used a transformer-based, fine-tuned T5 model for generating research highlights and secured the 10th position in the SciHigh track. The model was able to convert lengthy scientific abstracts into concise and well-structured highlights, demonstrating the effectiveness of T5 as a strong baseline for this task.

However, the approach has certain limitations, including restricted dataset coverage and the computational overhead of fine-tuning T5 for domain-specific summarization. Future improvements may include expanding the dataset to cover more research areas, enhancing factual consistency, and integrating human feedback or richer semantic evaluation metrics to further refine the quality of generated highlights.

Declaration on Generative AI

The author(s) used Generative AI tools solely for grammar and spelling checks. All experimental work and analyses were carried out independently by the author(s).

References

[1] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[2] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[3] R. Mihalcea, P. Tarau, TextRank: Bringing order into text, in: D. Lin, D. Wu (Eds.), Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 404–411. URL: https://aclanthology.org/W04-3252/.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020).

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[6] 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, ICML'20, JMLR.org, 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[8] Y. Liu, M. Lapata, Text summarization with pretrained encoders, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3730–3740. URL: https://aclanthology.org/D19-1387/. doi:10.18653/v1/D19-1387.

[9] T. Bao, H. Zhang, C. Zhang, Enhancing abstractive summarization of scientific papers using structure information, Expert Systems with Applications 261 (2025) 125529. URL: https://www.sciencedirect.com/science/article/pii/S0957417424023960. doi:https://doi.org/10.1016/j.eswa.2024.125529.

[10] H. Jing, K. R. McKeown, Cut and paste based text summarization, in: 1st Meeting of the North American Chapter of the Association for Computational Linguistics, 2000. URL: https://aclanthology.org/A00-2024/.

[11] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, ArXiv abs/1409.3215 (2014). URL: https://api.semanticscholar.org/CorpusID:7961699.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: https://arxiv.org/pdf/1706.03762.pdf.

[13] R. Rezapour, Y. Ge, K. Han, R. Jeong, J. Diesner, Two-stage graph-augmented summarization of scientific documents, in: L. Peled-Cohen, N. Calderon, S. Lissak, R. Reichart (Eds.), Proceedings of the 1st Workshop on NLP for Science (NLP4Science), Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 36–46. URL: https://aclanthology.org/2024.nlp4science-1.5/. doi:10.18653/v1/2024.nlp4science-1.5.

[14] T. Rehman, S. Das, D. K. Sanyal, S. Chattopadhyay, Abstractive text summarization using attentive

gru based encoder-decoder, in: Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2021, Springer, 2022, pp. 687–695.

[15] T. Rehman, S. Das, D. K. Sanyal, S. Chattopadhyay, An analysis of abstractive text summarization using pre-trained models, in: Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2021, Springer, 2022, pp. 253–264.

[16] E. Collins, I. Augenstein, S. Riedel, A supervised approach to extractive summarisation of scientific papers, in: Proc. 21st Conf. on Computational Natural Language Learning (CoNLL 2017), ACL, Vancouver, Canada, 2017, pp. 195–205.

[17] L. Cagliero, M. La Quatra, Extracting highlights of scientific articles: A supervised summarization approach, Expert Systems with Applications 160 (2020) 113659. URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417420304838.

[18] T. Rehman, D. K. Sanyal, S. Chattopadhyay, P. K. Bhowmick, P. P. Das, Automatic generation of research highlights from scientific, in: 2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE'21), collocated with JCDL'21, 2021.

[19] T. Rehman, D. K. Sanyal, P. Majumder, S. Chattopadhyay, Named entity recognition based automatic generation of research highlights, in: Proceedings of the Third Workshop on Scholarly Document Processing (SDP 2022) collocated with COLING 2022, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 163–169. URL: https://aclanthology.org/2022.sdp-1.18.

[20] T. Rehman, D. K. Sanyal, S. Chattopadhyay, Research highlight generation with elmo contextual embeddings, Scalable Computing: Practice and Experience 24 (2023) 181–190.

[21] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[22] 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. doi:10.1109/ACCESS.2023.3292300.