*ICIC Express Letters*
Volume **18**, Number **2**, February **2024**
 ISSN 1881-803X
pp. **193–200**

# EMBRACING TEXT SUMMARIZATION IN EDUCATION USING TRANSFORMER MODEL

Ilvico Sonata* and Yaya Heryadi

Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
Bina Nusantara University
Jl K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
yayaheryadi@binus.edu
*Corresponding author: ilvico@binus.ac.id

Received February 2023; accepted May 2023

Abstract. *The number of books and articles available in digital libraries and literature, particularly on education-related topics, keeps increasing in the past decade. Such fast growing number of texts is impossible to read one by one by anyone in order to comprehend the main points of the texts. With such a big amount of texts as references in the digital space, an efficient text summarization is very crucial in order to automatically shorten long piece of texts and produce accurate summaries that can fluently pass the main points from the texts. Text summarization task is a problem that aims to create text summary or abbreviation by shortening long pieces of text while preserving the main points drawn in the text. The method to address this problem is very instrumental in many domains including education field. This paper presents an exploratory result in developing a text summarizer model using the Transformer model to summarize various textual document related to education in Indonesia. The empiric results showed that the proposed model can achieve the precision value of ROUGE-1 as 0.778, ROUGE-2 as 0.762 and ROUGE-L as 0.814, the recall value of ROUGE-1 as 0.892, ROUGE-2 as 0.857 and ROUGE-L as 0.895, and the F-Score value of ROUGE-1 as 0.875, ROUGE-2 as 0.749 and ROUGE-L as 0.886. The testing results show that the accuracy produced by the Transformer model is quite promising and produces higher accuracy compared to Generative Adversarial Networks (GAN) based model as baseline.*
**Keywords:** Education literature, Natural language processing, ROUGE, Text summarization, Transformer

1. **Introduction.** Text summarization is a task that aims to create text summary or abbreviation by shortening long pieces of text which preserves the main points drawn in the text [1]. Text summarization is an interesting research topic in Natural Language Processing (NLP) with wide potential applications including, but are not limited to, education field. In a teaching and learning process, it is very common that teachers need to sort a large number of available supporting teaching materials to determine which articles were really needed and relevant to be assigned to students as reading materials. Hence, text summarization plays an important role in speeding up the process of selecting articles [1] or creating text summary of an article that contains key and important information [2].

In the past decade, text summarization has raised research interest from researchers in many research fields resulted in a vast number of proposed methods available in literature. Several methods are widely used to create text summarization models, including machine learning models such as Support Vector Machine (SVM) and Naïve Bayes [3] and deep learning models such as Recurrent Neural Networks (RNN) [4], Long Short-Term Memory (LSTM) [5,6], GAN [7,8] and Transformer [9-12].

DOI: 10.24507/icicel.18.02.193

The Transformer model introduced by Vaswani et al. [13] has become a state-of-the-art model in NLP. Several studies reported by Luo et al. [14] and Syed et al. [15] show some evidences that the Transformer model outperforms some previous models such as RNN and LSTM in terms of accuracy. However, comparisons between Transformer and GAN models have so far not been much done.

In the past decade, automated text summarization methods have been applied to summarizing books, social media posts, sentiment analysis, news, email, legal documents, biomedical documents, and scientific papers [16,17]. In education field, text summarization is very helpful for learners who have a deficiency in reading ability mainly to improve reading ability to comprehend a long piece of text [18] and remember important information from the text [2]. Despite many studies on text summarization have been published, it appears that there are only a few publications on text summarization in the field of education especially using text in Indonesian language as input. For this reason, the objective of this study is to propose a Transformer model as a text summarizer model for Indonesian text in the field of education and compare the accuracy with the GAN model as the baseline model. With the development of this text summarization model, the search for digital literature in the field of education in Indonesian can be faster with more precise results according to the topic needed.

In the next section, the proposed method will be discussed including the framework, methods and architectural models used. The experimental results and findings of the proposed method will be discussed in the results and discussion section. The final section contains the conclusions of this research and further research plans to improve the accuracy of the developed model.

## 2. Proposed Method.

The text summarization method in this study is the model-based abstractive method to extract the essence of the input text [19]. In this study, the proposed text summarizer model is a Transformer model which was first introduced by Vaswani et al. [13] in 2017. The use of this method is in accordance with the benefits of text summarization in education as described by Benzer et al. [2] and Nandhini and Balasundaram [18].

### 2.1. Proposed framework.

The proposed framework in this study can be illustrated using Figure 1.

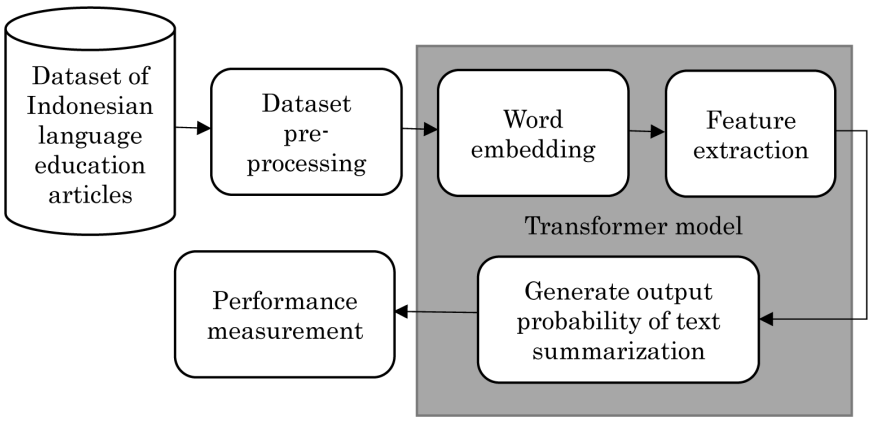


FIGURE 1. The proposed framework

The input dataset for this study is a labeled dataset $S = \{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$ which contains $n$-sample of text $x_i$ and its summary $t_i$ as the label. The data, which contains text about education in Indonesian languages, are collected from many sources. Summary of each data sample is prepared manually and served as the data label. Data preprocessing is implemented to the input data including tokenization, correcting word spelling,

converting sentence into lowercase, and removing some punctuation and non-ASCII characters [20]. Each token is then represented using word embedding.

The dataset is split randomly into training and validation dataset. A Transformer model is then trained supervisedly using an optimization algorithm and the labeled dataset as input. Performance of the trained model is measured using loss and ROUGE metrices.

2.2. **Transformer model architecture.** In general, a Transformer model is a deep learning model with deep structure architecture. Transformer model is a sequence-to-sequence deep learning model which has been used widely to address many tasks in NLP. Unlike many other models in neural network model family, a Transformer has attention mechanism to extract features needed in the text summarization process. The Transformer model architecture can be seen in Figure 2.

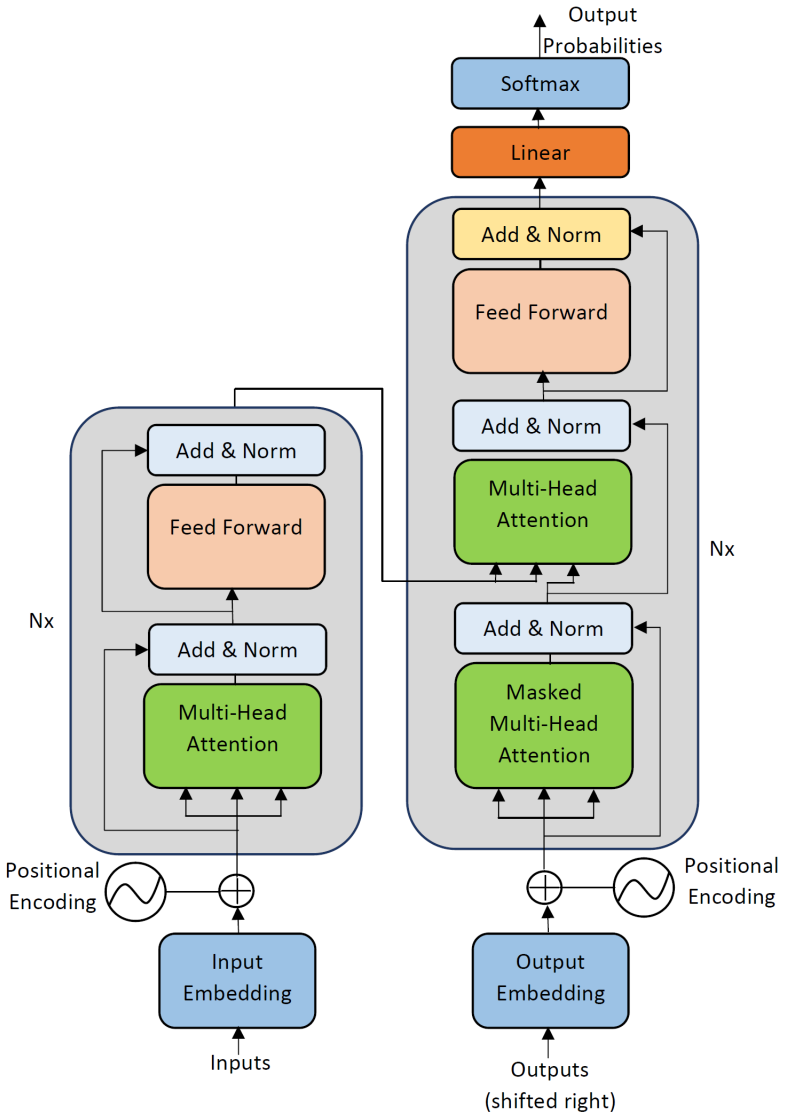


FIGURE 2. Transformer architecture

As can be seen from Figure 2, architecture of a Transformer consists of encoder and decoder parts. The input text as a sequence of words $S = \{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\}$ that enters the Transformer encoder is converted into a vector sequence through the embedding process. The order of words in the embedding process is very important so that they remain in position and the meaning of the sentences do not change. Positional encoding ensures that the word order stays in position [21]. The normalization layer placed

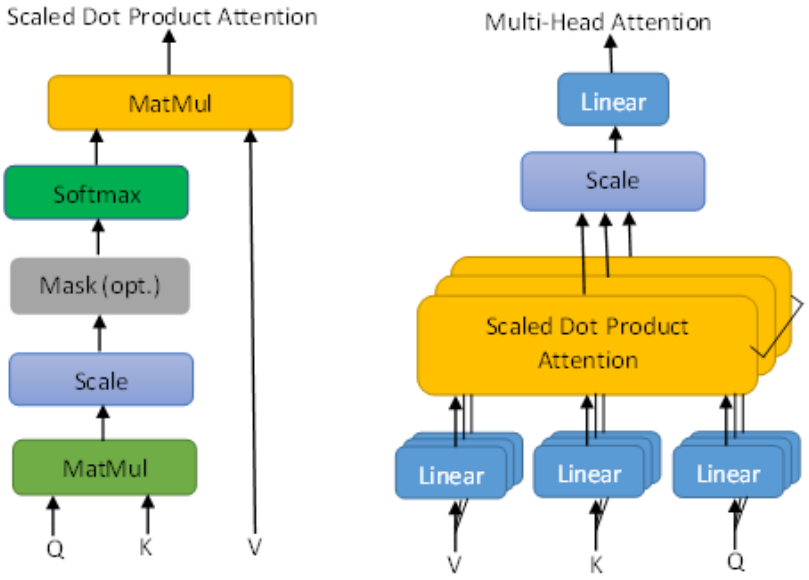


FIGURE 3. Attention mechanism

on each of the 2 sub layers of the Transformer architecture is useful for accelerating the training process through statistical processes [22].

The Transformer model uses an attention mechanism as shown in Figure 3 to extract features from the incoming word sequence. The incoming word order is weighed by Query ($Q$), Key ($K$) and Value ($V$) through the scaled dot product which is calculated using Equation (1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where $Q$, $K$, and $V$ are vectors and $d_k$ are the dimensions of vector $K$.

The dot product calculation is first performed for the $Q$ and $K$ vectors through a matrix multiplication process (MatMul). The dot product result is divided by the dimensions of the vector $K$ as part of the scaling process. The results of the scaling process are then applied to the softmax function to obtain relationships between words through a weighting process. Finally, the MatMul process is carried out from the softmax results with vector $V$ to produce the scaled dot product attention.

To improve the performance of this attention mechanism, a linear projection is performed on the vector dimensions $K$ ($d_k$), $Q$ ($d_q$), and $V$ ($d_v$) $h$ times to make the attention mechanism work in parallel. The output of multi-head attention is concatenated using Equations (2) and (3) as follows:

$$Multi\text{-}Head(Q, K, V) = Concat\left(head_1, head_2, \ldots, head_h\right)W^O, \tag{2}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{3}$$

where $W_i^Q \in \mathbb{R}^{d_{model} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$, $W$ is the metric weight for each $Q$, $K$ and $V$ obtained from the training results, $d_{model}$ is final output dimensions and $h$ can also be referred to as the number of heads to be used.

After going through the feature extraction process which is carried out with multi-head attention, the decoder generates a probability text summarization word sequence $y = (y_1, y_2, y_3, \ldots, y_m)$ which is processed by the Feed Forward Neural Network (FFNN). FFNN output can be calculated using Equation (4).

$$FFNN(x, t) = softmax\left(0, (x, t)W_1 + b_1\right)W_2 + b_2, \tag{4}$$

where $b_1$ and $b_2$ are bias.

The text summarization algorithm used in this study can be summarized as follows. First of all, the program will load the dataset for the training process, then preprocess the dataset and split the dataset for training and validation. Before the training process, hyperparameters are set for the Transformer model to be used. The Transformer model hyperparameter settings include learning rate, number of epochs for the training process, batch size, number of layers, number of heads and number of dense. The activation function used is softmax and the loss function used is Sparse Categorical Cross Entropy. Next, a training process of 30 epochs was carried out. The training process will produce a model with the best parameter weights to generate the text summarization. Algorithm 1 shows the algorithm used in the proposed Transformer model.

Algorithm 1. Text summarization model

**load** Indonesian language education articles dataset
**for each** Indonesian language education articles dataset
    tokenize, eliminate punctuation and non-ASCII, convert uppercase to lowercase
    word embedding
**set** train_size to 0.1
**set** validation_size to 0.9
**split** dataset into training and validation according to train_size and validation_size
**set** number of heads to 8, number of layers to 3, number of dense to 512
**set** batch size to 32, learning rate to 0.001, epoch to 30, activation function to softmax, loss function to Sparse Categorical Cross Entropy
**for each** word in dataset, **do** feature extraction using multi-head attention
    **for** epoch = 1: number of epochs
        **for** batch = 1: number of batches
            **Generate** another batch
            **Train** the model
            **Validate** the model
            **Backpropagate** the loss
**update** weight metric parameter
**generate** text summarization

3. **Result and Discussion.** The dataset used to train the Transformer model consists of 36,000 paragraphs of text about education written in Indonesian language. The input dataset is divided into training and validation dataset with proportion 90 : 10. The Transformer model is trained supervisedly using an optimization algorithm which is run in Google Colaboratory cloud computing system with GPU mode. The text summarizer model is trained using 30 epochs which takes 45 minutes training duration. The training result can be seen in Figure 4.

As can be seen in Figure 4, the training and validation loss curve by epoch is convergent to a small value which indicates that the model did not experience overfitting [23]. This condition indicates that the model can generalize to new data outside of the training data [24].

Following are two examples of input data and their respective summaries along with the summaries predicted using the proposed model:

1) Indonesian text as input: "*peran pendidikan dibutuhkan untuk mendidik peserta didik tidak hanya berupa bahan pembelajaran tetapi juga memastikan bahwa pendidik mampu meningkatkan karakter peserta didik untuk berperilaku baik*" (the role of education is needed to educate students not only in the form of learning materials but also to ensure that educators are able to improve the character of students to behave well);

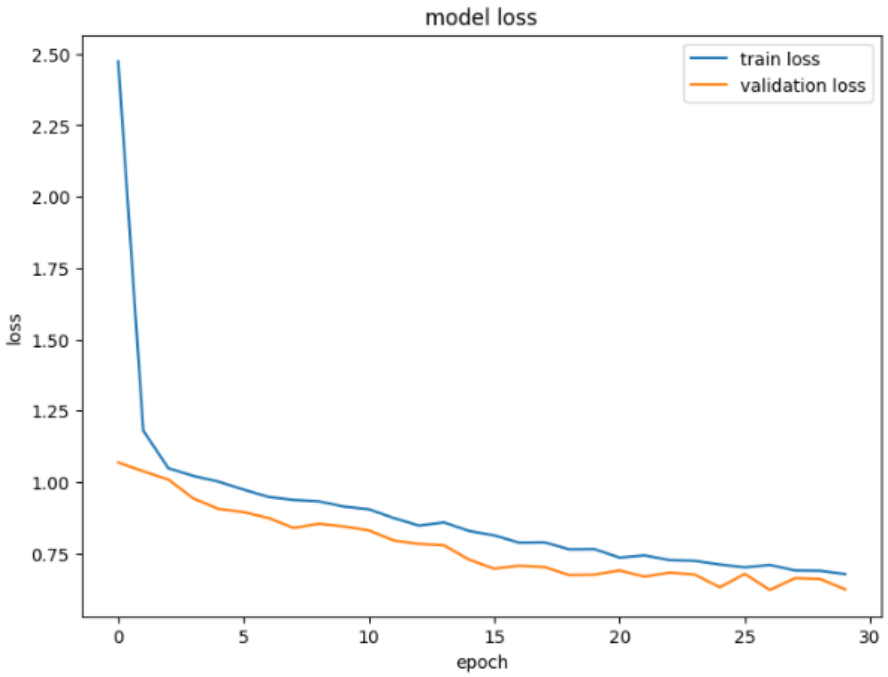


FIGURE 4. Training and validation loss result

2) Text summary as input: "*peran pendidik bukan hanya memberi bahan pembelajaran tetapi juga pengembangan karakter*" (the role of educators is not only to provide learning materials but also to develop character);
3) Predicted text summary: "*peran pendidik bukan hanya hanya memberi pembelajaran tetapi juga pendidikan karakter*".

Here is another example of predicted summary results:

1) Indonesian text as input: "*Pendidikan anak usia dini dimulai dengan mendidik akhlak luhur anak melalui kegiatan yang menyenangkan dengan peran orang tua di dalam keluarga*" (Early childhood education begins with educating children's noble character through fun activities with the role of parents in the family);
2) Text summary as input: "*Pendidikan anak usia dini dimulai dari keluarga*" (Early childhood education starts from the family);
3) Predicted text summary: "*Pendidikan anak usia dini dimulai dimulai keluarga*".

Following [25], the metric used to evaluate model performance in generating text summary is Recall-Oriented Understudy for Gisting Evaluation (ROUGE). ROUGE metric is generally used to evaluate the results of texts summarization by comparing the summary results generated from automated systems with standard human models. Several ROUGE metrics are as follows. ROUGE-1 is used to measure the similarity of one word. ROUGE-2 is used to measure the similarity of 2 words. ROUGE-L is used to measure the similarity of long word sequences. The ROUGE results from this experiment can be seen in Table 1.

TABLE 1. ROUGE results of the proposed model

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Precision | 0.778 | 0.762 | 0.814 |
| Recall | 0.892 | 0.857 | 0.895 |
| F-Score | 0.875 | 0.749 | 0.886 |

To further see the performance of the proposed model, the ROUGE results of the proposed model will be compared with the GAN model as proposed by Bhargava et al. [7] by using the same dataset as used in this proposed model. The reason for selecting the GAN model as the baseline model is because there are still few researchers who compare the performance of the Transformer model with the GAN model in text summarization compared to the RNN and LSTM models [15] where in general the Transformer model has better accuracy. In general, the GAN model as proposed by Bhargava et al. consists of a generator and a discriminator. The generator part is used to generate sample sentences and the discriminator part functions to determine whether the sample sentences are similar to the training data or not by using binary classification. Through this mechanism, the discriminator learns to determine summarization based on the dataset.

ROUGE results using the GAN model can be seen in Table 2.

Table 2. ROUGE results of the GAN model

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| Precision | 0.773 | 0.728 | 0.748 |
| Recall | 0.742 | 0.619 | 0.731 |
| F-Score | 0.771 | 0.698 | 0.791 |

4. **Conclusions.** From the experimental results using the same dataset, it appears that the Transformer model has better accuracy in terms of precision, recall and F-Score for ROUGE-1, ROUGE-2 and ROUGE-L than the GAN model proposed by Bhargava et al. [7] in generating Indonesian language text summarization for educational texts and articles. With these results, the Transformer model can be used to develop education, especially in Indonesia by quickly searching for appropriate education literature and helping learners understand articles by reading the essence displayed through a text summarization.

Further studies can be carried out by increasing the accuracy of this model by increasing the number of datasets for the latest articles in the field of education.

## REFERENCES

[1] S. A. Babar and P. D. Patil, Improving performance of text summarization, *Procedia Computer Science*, vol.46, pp.354-363, DOI: 10.1016/j.procs.2015.02.031, 2015.

[2] A. Benzer, A. Sefer, Z. Ören and S. Konuk, A student-focused study: Strategy of text summary writing and assessment rubric, *Eğitim ve Bilim*, vol.41, no.186, pp.163-183, DOI: 10.15390/EB.2016.4603, 2016.

[3] Rahul, S. Adhikar and Monika, NLP based machine learning approaches for text summarization, *Proc. of the 4th International Conference on Computing Methodologies and Communication (ICCMC 2020)*, pp.535-538, DOI: 10.1109/ICCMC48092.2020.ICCMC-00099, 2020.

[4] S. Alhojely and J. Kalita, Recent progress on text summarization, *2020 International Conference on Computational Science and Computational Intelligence (CSCI 2020)*, pp.1503-1509, DOI: 10.1109/CSCI51800.2020.00278, 2020.

[5] R. V. Saraswathi, R. V. Chunchu, S. Kunchala, M. Varun, T. Begari and S. Bodduru, A LSTM based deep learning model for text summarization, *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pp.1063-1068, 2022.

[6] S. Preethi, M. S. K. Shibi, S. Sheshan, R. K. Grace and M. S. Geetha, Abstractive summarizer using Bi-LSTM, *2022 International Conference on Edge Computing and Applications (ICECAA)*, pp.1605-1609, DOI: 10.1109/icecaa55415.2022.9936215, 2022.

[7] R. Bhargava, G. Sharma and Y. Sharma, Deep text summarization using generative adversarial networks in Indian languages, *Procedia Computer Science*, vol.167, no.2019, pp.147-153, DOI: 10.1016/j.procs.2020.03.192, 2020.

[8] J. Lin and S. Li, A generative adversarial constraint encoder-decoder model for the text summarization, *2022 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2022)*, pp.63-68, DOI: 10.1109/ISEEIE55684.2022.00018, 2022.

[9] A. Glazkova and D. Morozov, Applying transformer-based text summarization for keyphrase generation, *arXiv.org*, arXiv: 2209.03791, 2022.
[10] S. Porwal, L. Bewoor and V. Deshpande, Transformer based implementation for automatic book summarization, *International Journal of Intelligent Systems and Applications in Engineering*, pp.123-128, 2022.
[11] H. Chouikhi and M. Alsuhaibani, Deep transformer language models for Arabic text summarization: A comparison study, *Applied Sciences*, vol.12, no.23, DOI: 10.3390/app122311944, 2022.
[12] K. Wen and L. Zhou, Research on text summary generation based on bidirectional encoder representation from transformers, *2020 2nd International Conference on Information Technology and Computer Application (ITCA 2020)*, pp.317-321, DOI: 10.1109/ITCA52113.2020.00074, 2020.
[13] A. Vaswani et al., Attention is all you need, *Proc. of the 31st International Conference on Neural Information Processing Systems*, pp.6000-6010, 2017.
[14] T. Luo, K. Guo and H. Guo, Automatic text summarization based on transformer and switchable normalization, *2019 IEEE Intl. Conf. on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom 2019)*, pp.1606-1611, DOI: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00236, 2019.
[15] A. A. Syed, F. L. Gaol and T. Matsuo, A survey of the state-of-the-art models in neural abstractive text summarization, *IEEE Access*, vol.9, pp.13248-13265, DOI: 10.1109/ACCESS.2021.3052783, 2021.
[16] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan and M. M. Kabir, A survey of automatic text summarization: Progress, process and challenges, *IEEE Access*, vol.9, pp.156043-156070, DOI: 10.1109/ACCESS.2021.3129786, 2021.
[17] J. L. Espejel, *Automatic Abstractive Summarization of Long Medical Texts with Multi-Encoders Transformer and General-Domain Summary Evaluation with WikiSERA*, https://tel.archives-ouvertes.fr/tel-03376172/document, 2021.
[18] K. Nandhini and S. R. Balasundaram, Improving readability through extractive summarization for learners with reading difficulties, *Egyptian Informatics Journal*, vol.14, no.3, pp.195-204, DOI: 10.1016/j.eij.2013.09.001, 2013.
[19] R. Boorugu and G. Ramesh, A survey on NLP based text summarization for summarizing product reviews, *2020 2nd International Conference on Inventive Research in Computing Applications (ICIRCA 2020)*, pp.352-356, DOI: 10.1109/ICIRCA48905.2020.9183355, 2020.
[20] M. V. Wüthrich and M. Merz, *Statistical Foundations of Actuarial Learning and Its Applications*, Springer Cham, 2022.
[21] P. C. Chen, H. Tsai, S. Bhojanapalli, H. W. Chung, Y. W. Chang and C. S. Ferng, A simple and effective positional encoding for transformers, *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp.2974-2988, DOI: 10.18653/v1/2021.emnlp-main.236, 2021.
[22] R. Xiong et al., On layer normalization in the transformer architecture, *Proc. of the 37th International Conference on Machine Learning*, vol.PMLR 119, pp.10524-10533, 2020.
[23] J. Kolluri, V. K. Kotte, M. S. B. Phridviraj and S. Razia, Reducing overfitting problem in machine learning using novel L1/4 regularization method, *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp.934-938, 2020.
[24] C. Xu, P. Coen-Pirani and X. Jiang, Empirical study of overfitting in deep FNN prediction models for breast cancer metastasis, *arXiv.org*, arXiv: 2208.02150, 2022.
[25] C. Y. Lin, ROUGE: A package for automatic evaluation of summaries, *Text Summarization Branches Out (WAS 2004)*, no.1, pp.25-26, papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85, 2004.