Proceedings of 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS 2024)
IEEE Xplore Part Number: CFP24DJ3-ART; ISBN: 979-8-3503-7999-0

Text Summarization based on Feature Extraction using GloVe and B-GRU

R.RoselinKiruba
department of SOC
Vel Tech Rangarajan Dr.Sagunthala R & D Institue of Science and Technology
Avadi, Chennai
kirubaroselin@gmail.com

S. Sowmyayani
deptartment of CSE
St. Mary's College (Autonomous)
Thoothukudi, India
sowmyayani@gmail.com

S.Anitha
department of BCA
Govindammal Aditanar College of woman
Tiruchendur, India
anithajayakodi@gmail.com

J Kavitha
department. of SOC
Vel Tech Rangarajan Dr.Sagunthala R & D Institue of Science and Technology
Avadi, Chennai
vsskavitha@gmail.com

R.Preethi
department of english
Vel Tech Rangarajan Dr.Sagunthala R & D Institue of Science and Technology
Avadi, Chennai
rajampreethi@gmail.com

C.Saranya jothi
department of SOC
Vel Tech Rangarajan Dr.Sagunthala R & D Institue of Science and Technology
Avadi, Chennai
saranyajothi22@gmail.com

2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS) | 979-8-3503-7999-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSCSS60660.2024.10625311

***Abstract*—In the day-to-day life, huge amount of data needed Automated Text Summarization (ATS) methods to identify the useful information. The traditional summarization methods are very complex as it consumes more time for people. In order to solve this issue lot of sentence scoring or ranking methods for labelling the input text is introduced for the summary. Therefore, Feature Extraction (FE) with deep learning techniques are emerging as a solution. However, the Bi-Gated Recurrent Unit (B-GRU) has the limitation in missing the features. To update with a more appropriate solution, the novel approach is incorporated in FE with sentence similarity and word feature vectors separately using Global Vectors for Word Representation (GloVe). This approach also adds B-GRU with sliding windows for more FE with attention layers. This makes the ranking for the important sentence that is included in the summary. The experimental results are implemented in arXiv and PubMed datasets. The proposed method outperforms with the various several state-of-the-art models. This method gains the remarkable Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1, ROUGE-2 and Recall-Oriented Understudy for Gisting Evaluation Longest (ROUGE-L) scores in both the datasets. It shows the proposed method have the potential with the informative summaries as it has good precision, Recall and F-score.**

Keywords—Deep Learning, Feature Extraction, Text Summarization.

I. INTRODUCTION

Social media and online boards have arisen as the most extensively used platforms for people to talk about their experiences and gain knowledge. The struggle to gain appropriate information has grown nowadays due to the large information accessible online and published on websites. People find it tiresome to read abundant articles that contain redundant material these days. Therefore, it is vital to have an automatic summary system that may assist in swiftly finding the most essential and obvious information in a short duration. Automatic summarization is used for countless provinces such as news, webpages, search engines, and all kinds of online reviews.

In the study conducted by [1], the process of summarizing was abstracted as a difficulty of making complete inferences, to concurrently optimize three aspects: relevance, redundancy, and length. The scoring function of this research is resemblance to Maximal Marginal Relevance (MMR). Wang et al. [2] announced a Bayesian Sentence-based Topic Model (BSTM) that operates both term-sentence documents and term-sentence relationships for multi-document summarization. It applied probability distributions to signify the chance of selecting phrases based on subjects that offer a systematic approach for the task of summarization. Multi-document summarization is a multi-objective optimization problem. The optimization-based document summarizing model has yet to be carefully scrutinized. It demands the simultaneous optimization of many goal functions. An effective summary should expansively cover the main points offered in the documents. It should evade needless repetition and a continuous flow between phrases. Qumsiyeh et al. [3] presented a query-based summarizer to advance the web search engine results. It is commonly accredited that one of the most noteworthy applications of Natural Language Processing (NLP) is text summarization. The purpose of this is to alter one or more connected text documents into a more succinct version while maintaining the original content and meanings of the document as a whole. Input, methodology, language, generality and output are the categories that can be used as a way for summarization. Modaresi et al, [4] published a study that demonstrates the influence of using the query-based extractive summarizing approach for media monitoring and media reaction analysis.

Several deep learning-based text summarization techniques have increased researcher attention due to their accuracy. Several studies have explored the use of Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM) in the extractive summarization. Nallapati et al. [5] separates the summarization into binary classification with summary and non-summary category. The GRU-Recurrent Neural Network (RNN) model is develop to map the position of the phrases to reduce the redundancy. This can be used as the both abstractive and the extractive summary. It established that the resulting summarizing system strategy provides competitive results when applied to DailyMail and DUC 2002 datasets. The sentence and document extracted using

the encoder using Neural Extractive Document Summarization (NeuSum) developed by Zhou et al. [6] GRU-RNNs to score and select the sentences. In order to avoid repetition, the RNNs phrase extractor is used to select the sentence perfectly. Shi et al. [7] developed a novel method for extractive summarization method for capturing the salient using RNN-GRU. For extracting the sentence features Convolution Neural Network (CNN) is followed by B-GRU method is adopted. Similarly, in the paper [8] enhances the text Feature Extraction (FE) along with the emotions in the data. It also uses the CNN for feature extraction and for semantic learning purpose the B-GRU is used. Zogan et al. [9] also B-GRU for different purpose like understanding the sequence of sentence, finding the linguistic patterns. Sheher et al. [10] uses the power of BERT which is pre-trained on extensive self-supervised datasets. It is united with BiGRU which is a recurrent neutral network. This captures the dependencies within the text to extract salient information. This work comprises of BERT to get sentence embedding which are given into BiGRU network the collaboration of BERT ad BiGRU improvised the extractive text summarization.

This study comprises creating more effective models that can integrate external knowledge and contextual data, inventing faithful techniques for handling large amounts of data, and discovering novel methods for summarizing diverse forms of media, images, and videos. This research supports researchers who are concerned in this discipline and arouses supplementary progress. Although there are several methods proposed in the traditional text summarization technique is based on constraints and assumptions such as; precision, accuracy and F-score. The proposed work is used to overcome some of the drawbacks addressed above by devising an existing deep learning techniques.

The innovation of the work is summarized as the following:

1) FE with two sentence similarity and sentence with title sentence similarity is adopted.
2) FE with word vectors are properly used for identifying the sentence semantics.
3) B-GRU techniques for enhanced features like bidirectional and sliding window make this method to give good summarization results.

This research work is organized as follows: section 2 presents the overview of proposed work; section 3 presents the proposed data hiding algorithm followed by experimental results and conclusion in section 4.

II. Proposed Methodology

Text Summarization is a major concern in NLP, which exploit the useful information from the unstructured data to form the summarization. Given length document is first pre-processed and then each sentence and word features are extracted using GloVe process. While extracting the sentence features, both the title similairty sentence and two sentence similarity are concentrated. Finally, the B-GRU for forward and backward GRU with attention layers. Finally, the approporiate text are label and ranked for optimal summarization is shown in Fig. 1. The details of explanation of each module are explained in the following subsections.

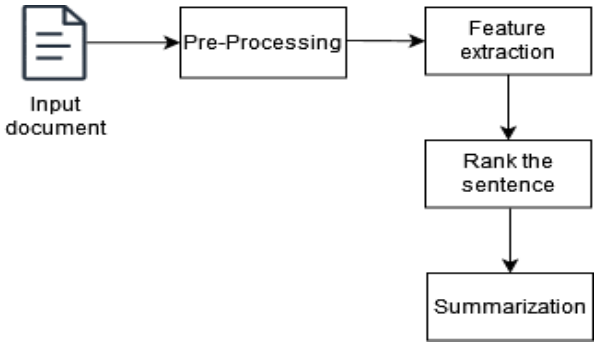

Fig. 1. Structure of the proposed method

A. Pre-processing

Pre-processing is the technique is used to convert the document into efficient format by removing the unwanted data. This includes data aligning, tokenization, terminate word deletion, stemming, lemmatization, regulate text word and remove redundancy data is shown in Fig. 2.

i) **Text alignment:** The process of altering the text contains the following process they are:

- Making the words into one case (Ex. Eliminated as different words “Apple” and “apple”).
- Delete all the punctuation,
- Variation words expansion (Ex. “can’t” word is changed to “cannot”).
- Formalize the numbers to characters thereby reducing the symbols (Ex. “5” is converted to “five”).

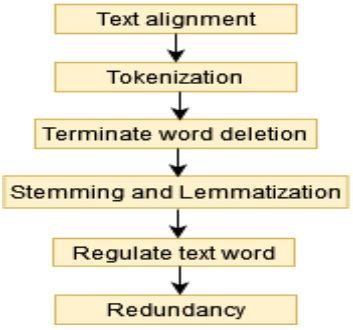

Fig. 2. Preprocessing of the proposed method

ii) **Tokenization:** Combined words are separated into words and sentence.
iii) **Terminate word deletion:** Remove the words like "the," "a," or "is" that don't contribute much to the core meaning of the text. This helps focus on important content.
iv) **Stemming and Lemmatization:** In stemming cutting words process in the suffixes is implemented (Ex. “playing” to “play)”. In lemmatization same syntactic meaning of words are replaced to form exact original word (Ex. “better” to “good”).
v) **Regulate text word:** It covert the abbreviations, typos and emojis to understand format.
vi) **Redundancy:** If the data is repeated many times it is removed.

B. Feature extraction

In Feature Extraction (FE), different features of sentence are extracted by using vectors. The selection of these features will give good quality to the summarization. In addition, the statistical and semantic features are included. The selective feature factors such as similarity of the title and original document, each sentence similarity and Glove standard for word representation is illustrated in Fig. 3. The FE used in the proposed approach are explained in the following steps.

i) Similarity of the title and original document

The similarity between the tile T and the sentence S are identified for each sentence. The sentence which match with the priority of the all title indicate the subject of the document. The resultant summarization should base on this observation. The degree of similarity of the title S_t and original document is given in Eq. (1).

$$S_t(S_i) = Similarity\ (S, T) \tag{1}$$

ii) Each sentence similarity

The resembles of one sentence from other sentence is calculated using cosine similarity using Eq. (2). Consider two sentence S_1 and S_2 is computed as

$$Similarity\ (S_1, S_2) = \frac{(S_1 \cdot S_2)}{\|S_1\|\|S_2\|} \tag{2}$$

Here, $S_1 \cdot S_2$ denotes dot product of the vectors and $\|S_1\|\|S_1\|$ and $\|S_2\|\|S_2\|$ represents the Euclidean values.

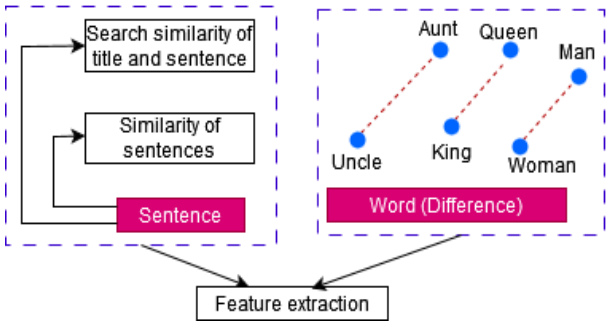

Fig. 3. Feature extraction

iii) Glove standard for word representation

The Global Vectors for Word Representation (GloVe). It is an unsupervised learning algorithm for generating the vector formation of words. It is pre-trained model for matching the word-to-word relationship of words. For example, the men and woman belongs to the human beings. It should identify how it differ from one another. A vector difference between two words vectors are generated to find the semantic relationship of words. The comparison and the contrast of the words are distinguish using GloVe. It learns the word vectors and the probability of repetition in the text is identified. Consider sample two words w_i and w_j and its vector can be given as v_i and v_j. The difference of vector v is given as in Eq. (3).

$$Difference(v) = v_i - v_j \tag{3}$$

The semantic correlation difference between the words with similar vector and dissimilar vector are calculated. The GloVe algorithm for word vectors is developed using Eq. (4).

$$GloVe\ (S_i) = \sum_{i,j=1}^{S} f(P_{ij}) \times (v_i \times v_j + b_i + b_j - \log P_{ij})^2 \tag{4}$$

Where, s is the size of the sentence, P_{ij} possibility of words, b_i and b_j base terms, f fixes the maximum weight for more repeated word and less weight for less repeated word. The sentence and word features are extracted in the form of vectors.

C. Automatic text summarization

In the previous subsection, we discussed how FE score generated for each sentence within the transcript. The method of Automatic Text Summarization (ATS) in NLP is used to train Bi-Gated Recurrent Unit (B-GRU) systems. The GRU train faster than the LSTM due to the advanced Recurrent Neural Network's (RNN) memory capacity using Fig. 4.

Step 1: Forward and backward process GRU: It contain two GRU in hidden state h_t which process the sentence features in the forward other in the backward direction. It is computed using the Eq. (5) and (6).

$$\vec{h}_t = GRU(\overrightarrow{v_t}, \overrightarrow{h_{t-1}}) \tag{5}$$

$$\vec{h}_t = GRU(\overleftarrow{v_t}, \overleftarrow{h_{t-1}}) \tag{6}$$

The above Equations (5) and (6) shows that the FE vectors with the time t. And the final combined B-GRU is given in the square bracket using Eq. (7).

$$\vec{z}_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{7}$$

In B-GRU the FE are passed to the linear layer 1 with the help of tanh activation process using Eq. (8).

$$\vec{z} = \tanh\ (l(\overrightarrow{h_t}, \overleftarrow{h_t}) = d_0 \tag{8}$$

Where, d_0 is considered as the input for GRU.

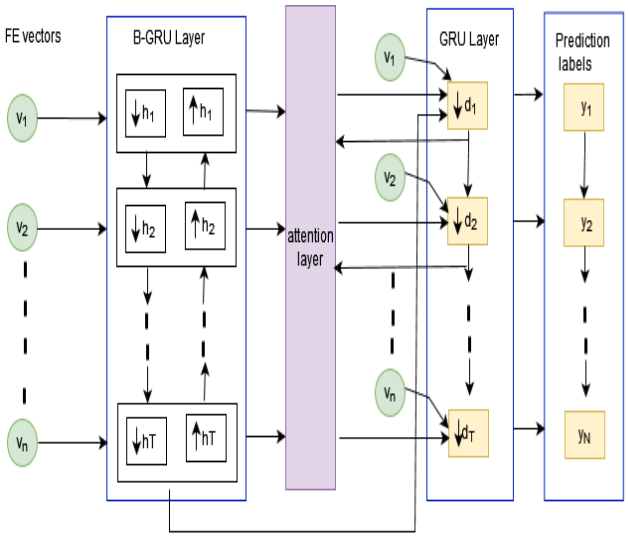

Fig. 4. Architecture of B-GRU

Step 2: Attention process: The attention layer is represented as a_t. It checks all the hidden state of the GRU and the final vector denotes the sentence with maximum priority (Eq. (9)).

$$\vec{a}_t = u \times E_t \tag{9}$$

Here, E represents the energy representation of GRU and previous hidden value d_{t-1}, u defined the weight of the hidden state, H is the hidden state. The energy expense is denoted as in Eq. (10).

$$E_t = \tanh\left(attn(d_{t-1}, H)\right) \tag{10}$$

Finally, the attention mechanism is passed to the softmax layer for normalization (i.e. value lies from zero and one). The weight source vector w_t is generated using the Eq. (11).

$$w_t = softmax(\vec{a}_t) \times H \tag{11}$$

Step 3: Label identification for summary: The previous hidden value d_{t-1}, feature vector v_i and the weight source vector w_t are given to the GRU is given to the Eq. (12).

$$d_t = BGRU(v_t, d_{t-1}, w_t) \tag{12}$$

For the sentence prediction, all the labels are given as the input to softmax and argmax function using Eq. (13) and (14).

$$\hat{y_t} = softmax(v_t, d_t, w_t) \tag{13}$$

$$y_t = argmax(\hat{y_t}) \tag{14}$$

The y_t gives the final prioritized sentence. The top sentences are stored as ranked and it is included in the summary.

III. EXPERIMENTAL RESULTS

In this section, the datasets, hyperparameters and evaluation metrics are discussed in detail.

A. Dataset

The two datasets namely ArXiv and Pubmed [11] as shown in Table I contain lengthy documents downloaded from arXiv.org and PubMed.com.

TABLE I. DATASET DESCRIPTION FOR TRAINING,VALIDATION, TESTING SET

Datasets	No. of Documents		
	Training set	Validation set	Testing set
ARXIV	203037	6436	6440
PUBMED	119224	6633	6658

TABLE II. DATASET DESCRIPTION FOR SENTENCES

Datasets	Average no of sentences	
	Doc	Summary
ARXIV	88	6.8
PUBMED	204	5.6

B. Hyper parameters

Hyper-parameters of GloVe and GRU are discussed. The various combination of batch size, learning rate, number of epochs, and optimizer makes the model to work effectively. For the GloVe, the batch size is 6 and initial learning rate is 0.00001 to minimize the memory during training. The epochs value is 8 and for tuning the model Adam optimization algorithm was used and the dropout value is 0.5. For GRU, similar Adam optimizer with the learning rate of 0.00001. The sliding window is set as 800 tokens wide. The window with maximum context is chosen.

C. Evaluation metrics

The evaluation of quality of summary in NLP is measured using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). The Recall-Oriented Understudy for Gisting Evaluation Longest (ROUGE-L). The ROGUE-L is used to find the Longest Common Subsequence (LCS) between O and R. The precision, recall and F_1 score are measured using ROGUE-L using Eq. (15)-(17). It is further divided into unigrams (ROGUE-1) and Bigrams (ROGUE-2).

$$ROUGE - L\ precision = \frac{LCS(O,R)}{Count(O)} \tag{15}$$

$$ROUGE - L\ Recall = \frac{LCS(O,R)}{Count\ (R)} \tag{16}$$

$$ROUGE - L\ F_1 score = \frac{2 \times ROUGE-L\ Precsion \times ROUGE-L\ Recall}{ROUGE-L\ Precsion + ROUGE-L\ Recall} \tag{17}$$

Where, $LCS(O,R)$ is the length measured for O and R, $count(o)$ and $count\ (R)$ total count of words in O and R.

TABLE III. MODEL COMPARISON USING ROUGE METRICS WITH ARXIV DATABASE

Models	arXiv		
	R_1	R_2	RL
Lead	34.1	8.96	21.2

LexRank	33.9	23.5	36.9
Match-sum	40.6	13	32.6
Seq2seq-local and global	43.6	17.4	29.1
BERTSUMEXT	41.4	14	35.2
BERTSUMEXT with sliding window	43.1	15.5	36
GRU with sliding window (Proposed method)	46.8	19.6	34.78

TABLE IV. MODEL COMPARISON USING ROUGE METRICS WITH PUBMED DATABASE

Models	Pubmed		
	R_1	R_2	RL
Lead	32.6	13	24.3
LexRank	56	26.7	40.4
Match-sum	41.2	14.9	36.8
Seq2seq-local and global	43.6	17.4	29.1
BERTSUMEXT	41.8	15	37
BERTSUMEXT with sliding window	44.9	20.2	39.4
GRU with sliding window (Proposed method)	47.1	21.5	39.8

From Table 3 and 4 it is notices that the proposed method had good performance in task summarization lengthy documents which is evaluated by ROUGE-1, and ROGUE-L. The proposed method is compared with the Bidirectional Encoder Representations from Transformers (BERT) which includes BERTSUMEXT [12, 13], which uses truncation method for to first 800 tokens in the document and it proves to be less efficient due to issue in handling long documents. Notably, the sliding window method of BERTSUMEXT, which align the documents effectively. The GRU method with the sliding uses B-GRU with two directions captures all the document using rank method which outperforms in producing effective summaries of long documents.

TABLE V. ROUGE-1 ON 3 RANDOMLY SELECTED DOCUMENTS FROM PUBMED DATASE

Document	Method	Pubmed (R1)		
		Precision	Recall	F-score
Doc-1	GRU with sliding window (Proposed method)	0.65	0.87	0.75
Doc-2		0.50	0.83	0.62
Doc-3		0.55	0.78	0.60
Doc-1	BERTSUMEXT with sliding window	0.63	0.85	0.74
Doc-2		0.49	0.82	0.61
Doc-3		0.53	0.77	0.59
Doc-1	BERTSUMEXT	0.45	0.78	0.57
Doc-2		0.47	0.72	0.49
Doc-3		0.42	0.57	0.43

In order to find the performance of the proposed model with the metric precision, recall and F-score with sample random documents form the PubMed dataset (ROUGE-1 and ROUGE-L) was projected in Table 5 and Figure 5, 6 and 7. The results show that the proposed method outperforms in all the metrics. The recall and precision score of the proposed method is high indicates the most relevant information are captured from the original document. Furthermore, the high F-score of the proposed method indicates the superior performance.

Fig. 5. Percision value for PUBMED (ROUGE-L)

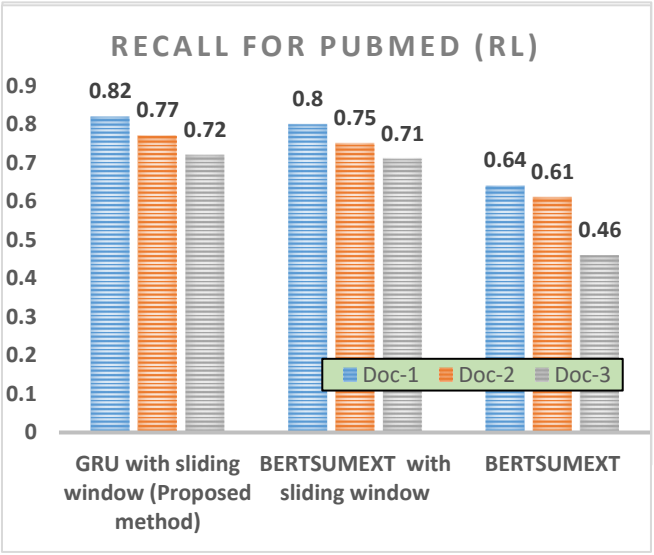

Fig. 6. Recall value for PUBMED (ROUGE-L)

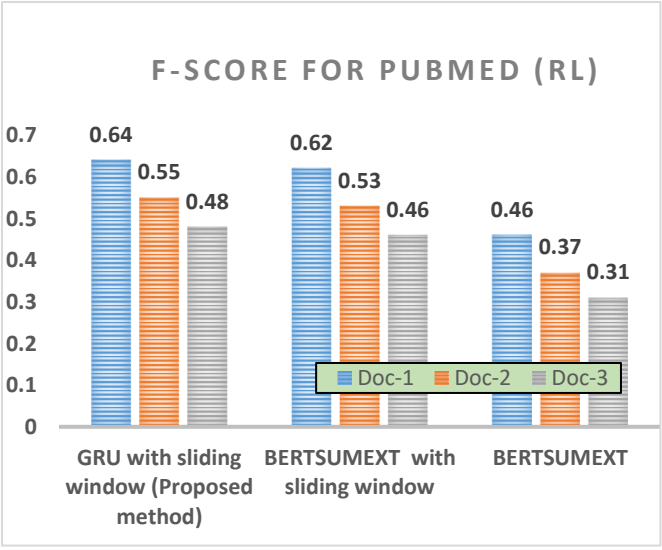

Fig. 7. F-score value for PUBMED (ROUGE-L)

IV. CONCLUSION

This paper proposes the sentence similarity, FE and B-GRU for ATS. In B-GRU the bidirectional FE are considered to extract the more sentence ranking. The two datasets are involved namely arXiv and PubMed for evaluating the performance. The highest accuracy is gained by the ROUGE-1 with a rise of 46.8 (arXiv) and 47.1 (PubMed), ROUGE-2 with a value of 19.6 (arXiv) and 21.5 (PubMed). The ROGUE-L has the value of 34.78 (arXiv) and 39.8 (PubMed) which has the highest rate value compared with the existing methods. Due to the specific FE technique and B-GRU for both the direction extraction along the maximum sliding window is utilized to increase the ROUGE-1 (precision (0.65), recall (0.87) and F-score (0.87) performance of the Document 1. Similarly, ROUGE-L (precision (0.60), recall (0.82) and F-score (0.64) performance of the Document 1 has the good results compared with the other random documents (Document 2 and Document 3). From this, it is validated that the proposed approach has the improved summarization from the base models such as BERTSUMEXT and BERTSUMEXT with sliding window techniques. For future work, study on the optimization algorithm can be explored to make the speedy text summarization.

REFERENCES

[1] McDonald, R., “A study of global inference algorithms in multi-document summarization”, In Proceedings of 29th European conference on IR research, LNCS, vol. 4425, Rome, Italy, pp. 557–564, 2007.

[2] Wang, D., Li, T., Zhu, S., & Ding, C, “Multi-document summarization using sentence-based topic models”, In Proceedings of the ACL-IJCNLP 2009 conference short papers, Singapore, pp. 297–300, 2009.

[3] R. Qumsiyeh and Y.-K. Ng, “Searching Web documents using a summarization approach," Int. J. Web Inf. Syst., vol. 12, no. 1, pp. 83101, Apr. 2016.

[4] P. Modaresi, P. Gross, S. Sedrodi, M. Eckhof, and S. Conrad, “On (commercial) benets of automatic text summarization systems in the news domain: A case of media monitoring and media response analysis," 2017.

[5] Nallapati, R., Zhou, B., Ma, M., “Classify or select: Neural architectures for extractive document summarization”, arXiv preprint arXiv:1611.04244, 2016. https://doi.org/10.48550/arXiv.1611.04244.

[6] Zhou, Q., Yang, N., Wei, F., Huang, S., Zhou, M., Zhao, T., “Neural document summarization by jointly learning to score and select sentences”, arXiv preprint, arXiv:1807.02305, 2018. https://doi.org/10.48550/arXiv.1807.02305.

[7] Shi, J., Liang, C., Hou, L., Li, J., Liu, Z., Zhang, H., 2019, “Deepchannel: Salience estimation by contrastive learning for extractive document summarization”, In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6999–7006. https://doi.org/10.1609/aaai.v33i01.33016999.

[8] Cheng, Y., Yao, L., Zhang, G., Tang, T., Xiang, G., Chen, H., Feng, Y., Cai, Z., “Text sentiment orientation analysis of multi-channels cnn and bigru based on attention mechanism”,. J. Comput. Res. Dev 57, 2583–2595, 2020.

[9] Zogan, H., Razzak, I., Jameel, S., Xu, G., “Depressionnet: A novel summarization boosted deep framework for depression detection on social media”, arXiv preprint arXiv:2105.10878, 2021. https://doi.org/10.48550/arXiv.2105.10878.

[10] Bano, Sheher, Shah Khalid, Nasser Mansoor Tairan, Habib Shah, and Hasan Ali Khattak. "Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach." *Journal of King Saud University-Computer and Information Sciences* 35, no. 9, pp. 101739, 2023.

[11] Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N., “A discourse-aware attention model for abstractive summarization of long documents”, 2018. arXiv preprint arXiv:1804.05685. https://doi.org/10.48550/arXiv.1804.05685.

[12] Liu, Y., “Fine-tune bert for extractive summarization”, arXiv preprint, arXiv:1903.10318, 2019.

[13] Kingma, D.P., Ba, J., “Adam: A method for stochastic optimization”, arXiv preprint arXiv:1412.6980, 2014. https://doi.org/10.48550/arXiv.1412.6980