Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TMRSG: TOPIC MODEL BASED RICH SEMANTIC GRAPH METHOD FOR ABSTRACTIVE MULTI-DOCUMENT SUMMARIZATION

Dr. K. ARUTCHELVAN, R. SENTHAMIZH SELVAN [2]

[1]Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India.

[2]Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India.

E-mail: [1]karutchelvan@yahoo.com, [2]mrsenthamizh@hotmail.com

ABSTRACT

Multi-Document Summarization (MDS) has gained more popularity among the industrialists and researchers in recent days. Extractive MDS simply extracts the important contents from multi-documents and gives a summary based on required length. Abstractive MDS provides summary based on the important of words that are presented across various documents. This research work is mainly focused on providing abstractive MDS using rich semantic graph-based methodology and topic modelling. The proposed approach generates summary by using graph-relations across multiple documents based on the relevant topics. The proposed approach is build using the centrality node ranking technique. The weighted graph ranking technique is applied to obtain the sequence of the sentences. The summary is generated using the highest rank scores of the sentences. The proposed technique is evaluated using the CNN/Daily Mail datasets.

Keywords: *Semantic Graph, Multi-Document Abstractive Summarization, Sentence Ranking, Similarity Measure, Topic Modelling*

1. INTRODUCTION

The advancement of technologies brought the users much closer with the internet application which is growing exponentially by seconds. Few decades before, people received the data from the media such as news channel, radio network, newspapers, etc. Nowadays, people are generating zettabytes of data through social-media application such as Facebook, LinkedIn, twitter, etc [1]. The data generates by the people are having voluminous of hidden information. Extracting the hidden knowledge from huge volume of data is a difficult task for researchers and industrialists. In the last two decades, Text mining (TM) is the dominant research field which is used to extract the insights from the large unstructured text data. In text mining, Natural Language Processing (NLP) is the core research field that has being exploring with multiple applications. Various applications such as sentiment analysis, recommendation systems, text summarization, text retrieval are using NLP to extract and understand the information that are available in the data. Text summarization (TS) is the task that is used to generate the summary from the huge collection of text documents. Text summarization is broadly applicable on single-document and multiple-documents [2].

Generating summary from the single document source is referred as single document summarization. Multiple-document summarization is the emerging research field where it generates summary from multiple sources with homogeneous information. Creating summary from the multiple documents is a difficult task. The two main categories are using to generate the summary that are as follows (i) extractive summarization and (ii) abstractive summarization [2]. Generating summary by choosing the vital contents from the large document is called as extractive summarization. The extractive summarization simply ranks the sentences and creates the summary. Abstractive summarization is to create summary from the sentences and produces abstractive summary. It

uses Natural Language Generation (NLG) concept to generate the summary. Lot of research works are being carried out in abstractive summarization. To avoid grammatical errors, extractive summarization is the best practice. To reduce the summary length abstractive summarization is the efficient methodology. Abstractive summarization also produces summary as human-generated summary. It can rephrase the sentences based on the advancement of techniques.

The proposed research work focuses on generating abstractive summary from multi-documents. For generating abstractive summary, it is essential to recognize the connection between the sentences. To find the connection among the sentences, most of the research work have been carried out in building semantic graph between the sentences. Rich semantic graph can be addresses with the important phrases and named entities of the document. The development of representation learning in NLP provides innovative approach on abstractive multiple document summarization [3]. In multiple homogeneous documents, the information is related to one another based on topics. Lot of research work have been carried in topic modelling. It is a key role in identifying the importance of topics and helps to generate the extractive or abstractive summary.

Major contributions of the proposed research article are as follows:

(i) The importance of topic modelling for text summarization is addressed in this article

(ii) The rich semantic graph is build using the LDA topic modelling

(iii) The topic for generating summary is identified using the named entity recognition technique

(iv) Predicate argument structure (PAS) is used to connect the semantics present in multi-document.

This research paper is organised as follows: Brief overview of topic modelling in abstractive summarization and graph-based techniques are discussed in the background study followed by the introduction. The proposed TMRSG (Topic Modelling based Rich Semantic Graph) method is presented in section 3. Outcome of the proposed work is discussed in the section 4. This research article is concluded in the section 5 with future directions.

2. BACKGROUND STUDY

In document modelling, topic model is being used little by little for generating the abstractive summary. The existing research articles [4, 5, 6] have been used the topic distributions as one of the features in their study. In the abstractive summarization process, topic extraction is a technique that is used in the pipeline. The topics are extracted using the various popular methods such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Probabilistic LSA (PLSA), and Gensim. LDA is the most popular method that is being used to extract the important keyword or topic from the source document. In the previous studies, most of the research work are carried out for the extractive summarization. In a work [7] the authors have used latent topic modelling to identify the important sentences. The authors built the cross-document relations using the Neural Topic Model (NTM) technique. In recent years, the DL (Deep Learning) models are applied to generate the abstractive summary [8].

Due to the lack of techniques most of the work has been achieved better ROUGE scores for extractive text summarization. The graph-based technique is being applied to produce effective summary [9, 10]. In the text summarization, abstractive summarization is a major challenging task whereas generating abstractive summary from multi-document summarization is a difficult process. Recent research articles are shown the interests of researchers in generating abstractive summary from multiple documents [11]. The abstractive summarization is generated using the graph-attention networks [11], categorical graph networks [18], hierarchical networks [12] and graph neural networks [13].

LDA was founded in [19-21]. It's a probabilistic generative model. Topic discovery is possible with LDA in natural language processing. The LDA's main concept is that texts are shown as random mixes on latent subjects, each of which is a word distribution. A group of documents has a probabilistic distribution of themes, meaning that some topics are more likely to be included in a given document than others. Within a subject, each term has its own probability distribution. That example, certain words are used far more frequently in a topic than others. In LDA, both sets of probabilities are generated using Bayesian techniques and the expectation maximization algorithm in the training phase. LDA has been

employed by certain researchers to perform MDS. Methods of extractive summarization based on LDA.

Avoiding the above draw backs of multi summarization problem by using the new method. The importance of topic modelling for text summarization is addressed in this article. The rich semantic graph is build using the LDA topic modelling. The topic for generating summary is identified using the named entity recognition technique. Predicate argument structure (PAS) is used to connect the semantics present in multi-document.

3. PROPOSED WORK: TMRSG

Extracting summary from the multiple documents can be achieved easily using the extractive text summarization (ETS). The ETS uses simple sentence ranking technique to find the important sentences in the multiple documents. It also uses the sentence similarity measures to reduce the redundant information [2]. Meanwhile, generating the abstractive summary is the difficult task. The topic modelling and rich semantic graph methods are used in the proposed work to generate the abstractive summary. The architecture of the proposed work is presented in the figure 1. The proposed work has three important components that are given below. The components are explained with definitions and algorithm.

(i) Preprocessing
(ii) Topic modeler
(iii) Rich semantic graph.

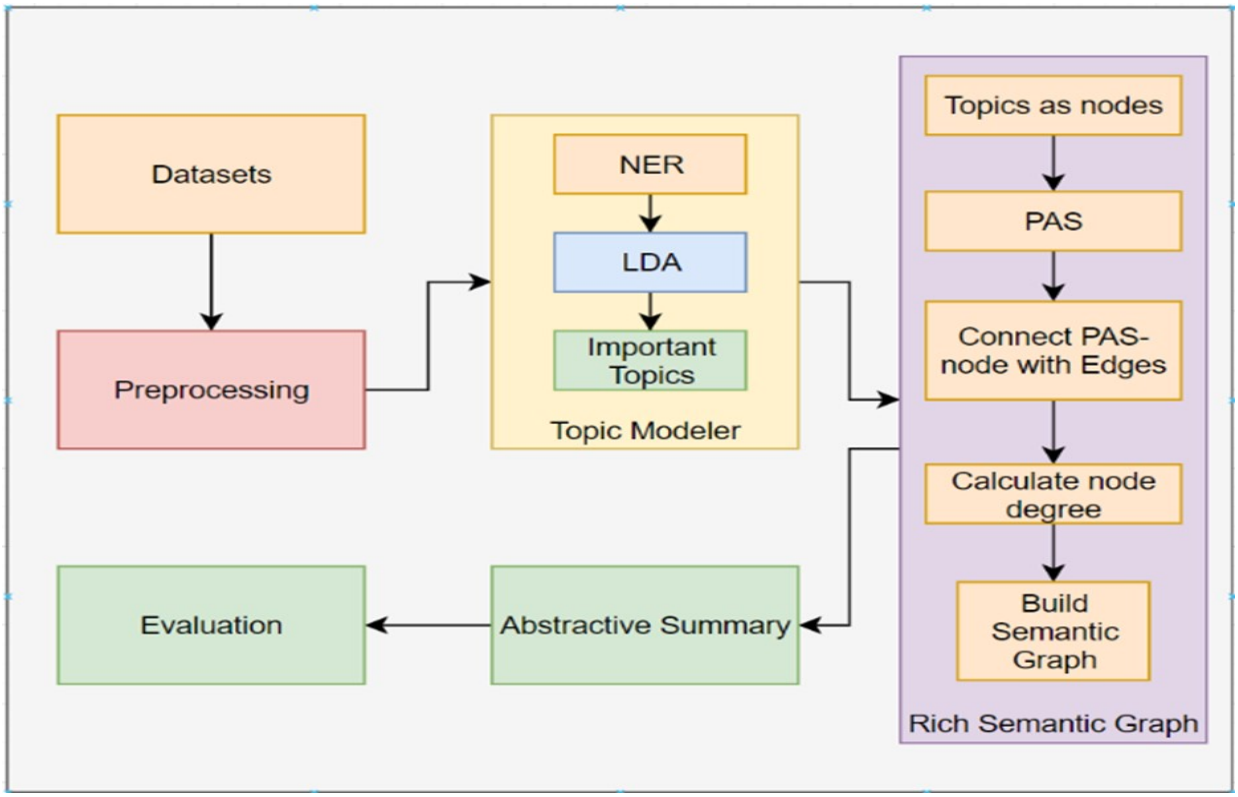

Figure 1: Research Flow Process

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Preprocessing

In the text mining process, data preprocessing is the important task to clean the given information. This component helps to filter the noisy words from the documents. Generally, in the text mining, the stop words are removed, but, in this work, the stop words are not removed. In generating summary, the stop words are valuable to discover the connectivity between the contents. The preprocessed texts are passed to the next component topic modeler.

Topic Modeler

Topic modelling is a tool that is used to find the hidden semantic structures in the text document. The proposed TMRSG algorithm uses, the most popular topic modelling Latent Dirichlet Allocation (LDA). The LDA is the generalized process of probabilistic latent semantic analysis (PLSA). The major assumptions that are considered by LDA is as follows: (i) all the documents are having a mixture of various topics (ii) all the topics are a combination of words or a topic is itself be a word. The topics in the documents are identified by the named entity recognition (NER) technique. The named entities such as person, location, organization, date and part-of-speech (POS) tagging. The named entities tagging is the fastest method to filter the important topics. This research work has used the Spacy 2.0 NER model to extract the entities. Spacy NER model is faster than any other NER model [14]. The identified named entities are tokenized with their generalized name and it is masked by the generalized name. The process of NER is given in the algorithm 1. In the proposed algorithm, it takes the input of sentences and tokenized the named entities across the multiple documents

The masked entities are used to find the classification of the topics in the given document. The topic classified based on the topics (i.e., entities) presented in the document. In multiple documents of homogeneous information, there is a highest probability for the same entities can be presented. The ranking of entities is provided by the LDA topic modelling. The NER generalization is given in the definition 1.

Definition 1: NER generalization and Topic Extraction

Assume classification C, set of topics T, set of documents as D and f(C) is the function that can be used to classify the documents based on the topics, where it comprises the following

$$f(C) = \begin{cases} d_i \in D, \text{ where } i - 1,2,3, \dots n-1, n, & i > 0 \\ t_j \in T, \text{where } j = 1,2,3, \dots, n-1, n, & j > 0 \cdots \end{cases}$$

The LDA model is represented in the figure 1.

Notations used:

- D: set of documents
- w: word in documents
- e: latent entity (topic) word in document
- θ: topic distribution in document
- x: parameter for topic distribution over the document
- y: parameter for word distribution over the topic
- v: word vector
- T: set of topics
- N: number of words present in the document

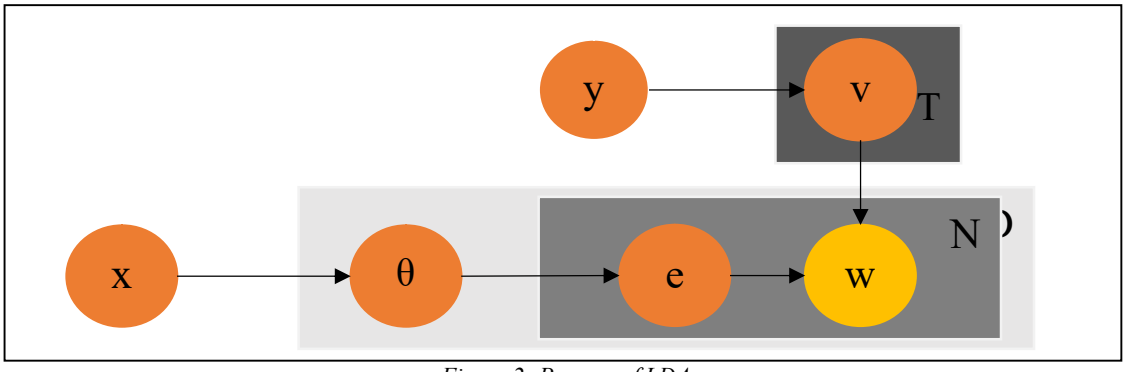

Figure 2: Process of LDA

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Algorithm 1: Topic Extraction

Input: set of documents

Output: Topics

1. T ← masked entities
2. D_{gen} ← empty
3. for d_i in D:
4. for d_j in D:
5. compute t_i in d_i
6. compute t_j in d_j
7. replace T with new masked entities
8. update D_{gen}
9. return D_{gen}

In the algorithm 1, T denotes the set of masked entities. For example, entities person is masked as person_1, person_2, … , person_n based on the number of different person presents in the document. Same entity can be presented in multiple documents. To compute the same entities, the step 3 – 7 matches the same entities (topics) in the document D and update the masked entities. D_{gen} denotes the generalized document. It contains the sentences with embedded words and masked topics. This helps to build the rich semantic graph.

Rich Semantic Graph

The topic modeler component produced the document with masked topics. It is known as; the sentences are having masked topics. Based on the sentence ranking, the relationship between the sentences is identified using the predicate argument structures (PAS). The POS tagging in the sentences are helpful to build the PAS. The topics are resembled as nodes and the relationship between the nodes is identified by the PAS.

Definition 2: Rich Semantic Graph

Assume G an unweighted-undirected graph that contains nodes and edges is denoted as G = (N, E) where N is the set of nodes and E is the set of edges (relationships between the sentences). Sentences in the documents are ranked using the degree of the vertices. The degree of vertices is calculated using the degree centrality given in the equation below:

$R = \sum_{i=0}^{D} d(m,n)$ where m and n are nodes.

Algorithm 2 represent the working step of the proposed TMRSG. The document with masked sentences is given as input. The D_{abs} denotes the abstractive summary that generates from the rich semantic graph method. The semantic similarity (SS) between the sentences is calculated using the cosine similarity measures. Step 2 – 6 provides the working procedure for linking the topics, identifies the semantic similarities and ranking the important sentence using degree centrality.

Algorithm 2: TMRSG

Input: Masked sentences

Output: Abstractive Summary

1. D_{abs} ← empty
2. for m_i in D_{gen}:
3. linked the masked entity with the predicate
4. calculate the semantic similarity between the predicates (p_i, p_j)

$$cos(p_i,p_j) = \frac{\sum_{k=1}^{n} p_i * p_j}{\sqrt{\sum_{k=1}^{n} p_i^2} X \sqrt{\sum_{k=1}^{n} p_j^2}}$$

5. find the node degree using degree centrality
6. update sentence ranking
7. D_{abs} ← sort the sentences based on rank
8. return D_{abs}

4.RESULTS AND DISCUSSIONS

The proposed TMRSG method is experimented with popular dataset CNN / Daily Mail. The dataset was published by Herman et.al [15] for the purpose of reading comprehension tasks and it is being in use for extractive and abstractive text summarization. The dataset can also be used for single-document and multi-document summarization, since, it comprises of 567 news articles. These news articles are generated under the relevant topic. The dataset contains total number of 3,11,971 long text of non-anonymous data. As it is discussed in the

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

previous section, the document is masked with the named entities (topics). The proposed TRMSG algorithm generates the abstractive summary based on the semantic relationships among the node with PAS. The experimental setup is implemented using python 3.9. Important libraries used for this research work is given in the table-1.

Table 1. Experimental Setup

Tools	Description
Spacy 2.0	Library for NER
networkx	Library for building semantic graph
numpy	Library for mathematical calculations
Word2vec	Library for word embedding
Pyrouge	Library for accuracy evaluation

The popular metrics ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score is used to evaluate the performance of proposed TMRSG method. Experiment is carried on different N-grams. The proposed method is compared with the baseline method TextRank [16] and LexRank [17]. The rouge score equation is given below,

$$ROUGE\ N(c,r) = \frac{\sum_{r_i \in r} \sum_{n\,grams \in r_i} Count(n\,grams, c)}{\sum_{r_i \in r} num\ Ngrams(r_i)}$$

where N is the grams (tokens/words), c and r denote candidate documents and reference documents respectively.

Table-2 represents the comparative results of the proposed work with baseline methods. Figure 3 depicts the chart representation of the table-2. From the results, proposed TMRSG method outperformed well than the existing baseline methods.

Table 2: Rouge Score Of Different Methods

Model	R-1	R-2	R-L	R-Average
LexRank	43.72	24.27	41.18	36.39
TextRank	41.85	22.45	39.47	34.59
TMRSG	44.71	25.42	42.92	37.68

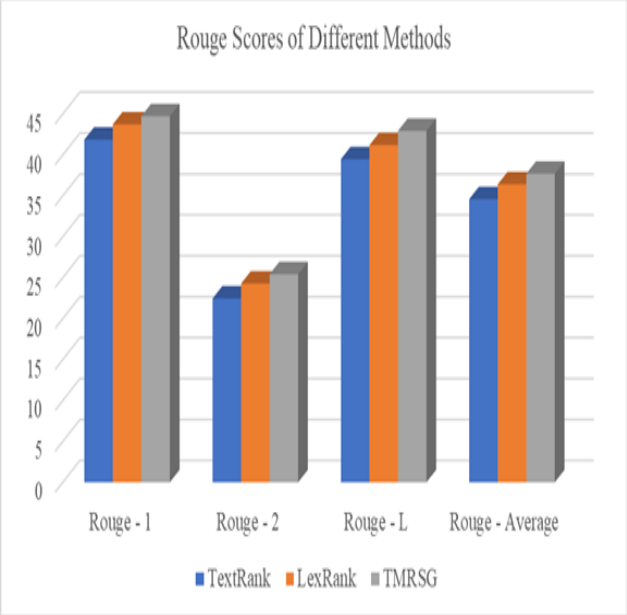

Figure 3: Rouge Scores Of Different Methods

5. CONCLUSION

In recent years, Multi-Document Summarization (MDS) has gained more popularity among the industrialists and researchers. The role of topic modelling in text summarization is discussed in this research articles. Named entity recognition provided by Spacy library finds the part-of-speech faster and accurately. The NER algorithm used to identify the topics present in the multiple documents. The predicate argument structure is used to build the rich semantic graph. The cosine similarity based semantic measure helps to reduce the redundant sentences in the documents. The experiment was conducted on the CNN/Daily Mail dataset. The proposed topic modelling based rich semantic graph method produced good accuracy compared to the baseline methods. In future, the proposed TMRSG method will be implemented using neural abstractive summarization technique.

REFERENCES:

[1] G. Vaitheeswaran, L. Arockiam, "Hybrid Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets", International Journal of Science, Engineering and Technology (IJCSET), June, 2016, Vol 6, Issue 6, 185-190.

[2] R. Senthamizh Selvan, Dr. K. Arutchelvan, "An Effective Approach for Abstractive Text Summarization using Semantic Graph Model", Annals of R.S.C.B, Vol. 25, Issue 4, 2021, April 2021, Pages. 13925 – 13933.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In 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), pages 4171–4186.

[4] Yang Wei. 2012. Document summarization method based on heterogeneous graph. In 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, pages 1285–1289.

[5] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1797–1807.

[6] Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020b. Friendly topic assistant for transformer based abstractive summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 485–497.

[7] Peng Cui, Le Hu, and Yuanchao Liu. 2020. Enhancing extractive text summarization with topic-aware graph neural networks. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5360–5371.

[8] Cui, Peng and Hu, Le, "Topic-Guided Abstractive Multi-Document Summarization", Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Nov, 2021, pages. 1463-1472, doi: 10.18653/v1/2021.findings-emnlp.126

[9] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In Proc. 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, July, pages 404–411.

[10] DanqingWang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020a. Heterogeneous graph neural networks for extractive document summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6209–6219.

[11] Atif Khan, Naomie Salim, Haleem Farman, Murad Khan, Bilal Jan, Awais Ahmad, Imran Ahmed, Anand Paul, "Abstractive Text Summarization based on Improved Semantic Graph Approach", International Journal of Parallel Programming, Springer Nature, 2018, https://doi.org/10.1007/s10766-018-0560-3

[12] Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5070–5081.

[13] Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6244–6254.

[14] Ramachandran, R., Arutchelvan, K. Named entity recognition on bio-medical literature documents using hybrid based approach. Journal of Ambient Intelligence Humanized Computing (2021). https://doi.org/10.1007/s12652-021-03078-z

[15] Hermann, K.M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; Blunsom, P. Teaching Machines to Read and Comprehend. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 1693–1701.

[16] Gunes Erkan and Dragomir R Radev, "Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization", Journal of artificial intelligence research, 2004, Vol. 22, Pages: 457–479.

[17] Rada Mihalcea and Paul Tarau, “Textrank: Bringing Order into Text”, In Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[18] R. Senthamizh Selvan, Dr. K. Arutchelvan, “Abstractive Summarization Using Categorical Graph Network”, Revista Geintec-Gestao Inovacao E Tecnologias, Vol. 11 No. 4 (2021), https://doi.org/10.47059/revistageintec.v11i4.2249.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[20] R. K. Roul, "Topic modeling combined with classification technique for extractive multi-document text summarization," Soft Computing, vol. 25, pp. 1113-1127, 2021.

[21] L. Na, L. Ming-xia, L. Ying, T. Xiao-jun, W. Hai-wen, and X. Peng, "Mixture of topic model for multi-document summarization," in Control and Decision Conference (2014 CCDC), The 26th Chinese, 2014, pp. 5168-5172