**A REVIEW PAPER ON TEXT SUMMARIZATION**

**Nikhil Chawla1, Palak Dhingra2, Pranav Goel3, Poonam Chaudhary4**

**Department of CSE, The NorthCap University, Gurugram, Haryana**

**1** [**nikhil18csu140@ncuindia.edu**](mailto:nikhil18csu140@ncuindia.edu)**, 2** [**palak18csu151@ncuindia.edu**](mailto:palak18csu151@ncuindia.edu)**, 3** [**pranav18csu158@ncuindia.edu**](mailto:pranav18csu158@ncuindia.edu)**, 4** [**poonamchaudhary@ncuindia.edu**](mailto:poonamchaudhary@ncuindia.edu)

**Abstract-** Text Summarization is a procedure in which important information is extracted from main text and represented in form of short summary. With the help of text summarizer we can alter the length of the data without losing its essence. The technique of automatic text summarization can be great assistance to save the time and extract the relevant information. There are two approaches for text summarization namely, Abstractive and Extractive. Extractive text summarization requires the drafting of words, expressions and sentences from the text to construct a new summary, while abstractive summarizer generates entirely new expressions and sentences to encapsulate the meaning of the original document.

***Keywords-*** *Natural language processing, text summarization, extractive summary, long short term memory, latent semantic analysis, latent dirichlet allocation.*

**Introduction-** Now a days, data is produced in large volumes with high velocity. In order to refer to such large amount of data a large portion of time is required. Even spending this much of time, the results are not relevant to search. Due to time consuming and noisy information, it may not efficient solution to go through such large amount of data manually. The technique of automatic text summarization can be great assistance to save the time and extract the relevant information. Text Summarazition comprise of two word ‘text’ means words and ‘summarization’ means stating meaningful information. Text Summarization is a procedure in which important information is extracted from main text and represented in form of short summary. With the help of text summarizer we can alter the length of the data without losing its essence. Text summarizer is actively used in day to day life like summary(from novels), headlines(from around the world), previews(from movies), bulletins(weather forcast). There are many text summarizers available online such as Microsoft News2, Google1, FreeSummarizer, WikiSummarizer & Text Compacter. There are two approaches for text summarization namely, Abstractive and Extractive. Extractive text summarization requires the drafting of words, expressions and sentences from the text to construct a new summary, while abstractive summarizer generates entirely new expressions and sentences to encapsulate the meaning of the original document. ting entirely new expressions and sentences to encapsulate the meaning of the text document. In other words, they translate and examine the text using advanced natural language techniques in order to produce a new summary that pass on the most critical information from the original text. This approach involves summarization based on deep learning and NLP, hence it is much harder than the extractive approach.

**1.Objective-** This paper consist of a text summarization model. Text Summarization is a procedure in which important information is extracted from main text and represented in form of short summary. There are two approaches for text summarization abstractive and extractive. The most common approach used is the extractive text summarization. In this paper, extractive approach of text summarization is used. The objective of this paper is to analyze and compare the supervised, unsupervised learning algorithm in text summarization. For unsupervised learning this paper has adopted LSA(latent semantic analysis) algorithm. To compare it with supervised learning algorithm, LSTM(long short term memory) has been implemented.

LSTM is an advance architecture of recurrent neural network (RNN) used in the domain of deep learning. It can be used to process both single point data as well as long sequences of data.It is appropriate for performing tasks such as time series prediction, connected handwriting recognition, speech recognition, etc. LSA is a technique in natural language processing, in particular distributional semantics, basically explores the relations between collection of documents and the words they contain by constructing a set of concepts related to the input documents and terms in them.

In section 2 we will talk about background study

**2.Background Study**

Moratanch et al.[1] have provided an exhaustive survey on abstractive text summarization methods. The two broad abstractive summarization methods are structured based approach and semantic based approach. The evaluation was done using various methods like human evaluation, rouge, F- score, recall, compression ratio, precision. They have concluded their paper by discussing the problem that it is impossible to build a standard model against which the results can be compared. A Survey of Text Summarization Extractive techniques have been presented by Gupta et al.[2]. They have discussed accuracy of the system which varies from 81% to 92 % after applying various evaluation methods like Intrinsic measure(human evaluation)and Extrinsic measure(task based performance). They have concluded their paper with the biggest challenge for text summarization is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user. Subramaniam et al.[3] presents a method for summarizing Hindi Text document by identifying substructures of graph that can extract meaningful sentences for generating a document summary. The approach used is Abstractive. They have optimized the summary by finding similarity among the sentences and merge the sentence which represented using Rich Semantic Sub graph which in turn produces a summarized text document. The parameters used by them are Sentence length, Average TF\_ISF (Term Frequency Inverse Sentence Frequency), Sentence position, Numerical Data. Chitrakalat et al.[4] has discussed that the pre-processed input document is used in the construction of bipartite graph. The bipartite graph captures the nested level of relationship between the sentences and concepts to ensure the highest level of efficiency in the extractive output summary. This can be used for single as well as multiple documents. They evaluated using different ROUGE methods like ROUGE-N, ROUGE-L, ROUGE-W, ROUGESU. They have concluded the paper by addressing the issues of abstractive summarization as there is no generalized framework, parsing and alignment. Reddy et al.[5] has proposed an approach which uses fuzzy classifier and deep learning algorithm. Fuzzy classifier produces score for each sentence and the deep learning (DL) also produces score for each sentence. The summarized text can be generated based on this hybrid score. By using this approach, they have achieved an average precision rate of 0.92 and average recall rate of 0.88 and the compression rate is 10% according to the experimental analysis. They have concluded that Hybrid Fuzzy algorithm is more effective that clustering algorithm. Banbyopadhyay et al.[6] have proposed multiple computational techniques like WordNet based, dictionary based, corpus based or generative approaches for generating SentiWordNet(s) for Indian languages. They have used two approaches k-means approach and Page Rank standard approach. They have concluded the results as precision-72.15%, recall-67.32% and f-score- 69.65 %. Kallimani et al.[7] have statistically analyze the adaptation of the methodology over multiple Indian languages and many document categories. They have provided an exhaustive survey using Abstractive method. They have used IE rules and class based templates in their model. They have evaluated the model using parameters with values as F score-0.815, Precision-0.8642, Recall-0.7973, Accuracy- 0.7217. Malik et al.[8] have discussed the idea to summarize Hindi text documents using sentence extraction method. They also used features such as statistical, linguistic feature and also used Genetic Algorithm. They have also concluded that sentence with similar feature is matched with subject of title. Bordag et al.[9] have provided a survey on Extractive method. The sentences are selected using an iterative extension of Page Rank calculation on a sentence similarity graph. Their evaluation is based on Rouge Score. They also concluded that while inspecting a number of summaries, similar sentences recurring often in texts are rarely selected by the Frank algorithm. Sony P et al.[10] have proposed a method in which sentence extraction is based on single document text summarization which produces a generic summary for a Malayalam document. The evaluation of the output is done by comparing the summarization outputs with manual summaries generated by human evaluators. They have used extractive method and more specifically TF-IDF techniques. They have also mentioned that there is no efficient WordNet for Malayalam, hence a synset for the corpus under consideration was use. They evaluated the model using compression ratio. Sinha et al.[11] used extractive text summarization. The main idea used was to break the entire text into pages. They have evaluated on the basis of ROUGE-1 and ROUGE-2. They assumed that summary length to be generated should be less than ‘page\_len’ which they will try to improve upon in the future. Song et al.[12] have pre-processed using stanford natural language process tool CoreNLP and ROUGE for evaluation. They have concluded their results on the datasets CNN and DailyMail showing their ATSDL framework outperforms the state-of-the-art models in terms of both semantics and syntactic structure. For building the model they concluded that determining semantic similarity between phrases is difficult. Lee et al.[13] build a model where, sentence-level attention is used to modulate the word-level attention such that words in less attended sentences are less likely to be generated. They have concluded that by training their model with the inconsistency loss and original losses of extractive and abstractive models, they achieve state-of-the art ROUGE scores. Liu et al.[14] describes the DISCOBERT model extracts sub-sentential discourse units as candidates for extractive selection on a finer granularity. They concluded that their proposed model outperforms state-of-the-art methods by a significant margin on popular summarization benchmarks compared to other BERT-base models. They used ROUGE-1, ROUGE-2 and ROUGE-L F-1 for evaluation. Lobo et al.[15] proposed a model which consist of reduction of a text document to generate a new form which conveys the key meaning of the contained text. They have mentioned that data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. They have also presented a combined approach to document and sentence clustering as an extractive technique of summarization. Rodriguez et al.[16] have used a Multi-Objective Artificial Bee Colony (MOABC) algorithm for implementing this task.They have used datasets from Document Understanding Conference (DUC) and model performances has been evaluated with ROUGE metrics. They have concluded that the results of the proposed approach shows important improvement, i.e., in average, 31.09% (8.43%) and 18.63% (6.09%) of improvement in ROUGE-2 (ROUGE-L). Abekawa et al.[17] have proposed a model which uses framework of reinforcement learning. They demonstrate that the method of reinforcement learning can be adapted to automatic summarization problems naturally and simply. They concluded that the results have revealed ASRL can search for sub-optimal solutions efficiently under conditions for effectively selecting features and the score function. Gupta et al.[18] have explored text summarization with an intermediate step of Abstract Meaning Representation (AMR). The pipeline proposed by them generates an AMR graph of an input story, through which it extracts a summary graph and finally, generate summary sentences from this summary graph. They concluded that the proposed method achieves state-of the-art results compared to the other text summarization routines based on AMR. Zhu et al.[19] have evaluated on the basis of 13,017 manually labeled sentences and statistically sampling of 3,456 extracted aspects, they confirm the high accuracy of the extraction method. They conclude that the composed CVE descriptions achieve high ROUGH-L (0.38). Ramakrishnana et al.[20] have developed an open source library, namely, Neural Abstractive Text Summarizer (NATS) tool kit, for the abstractive text summarization. They have performed experiments on CNN/Daily Mail dataset to examine the effectiveness of several different neural network components. They have concluded that they have implemented NATS on two models on the two recently released datasets, namely, Newsroom and Bytecup.

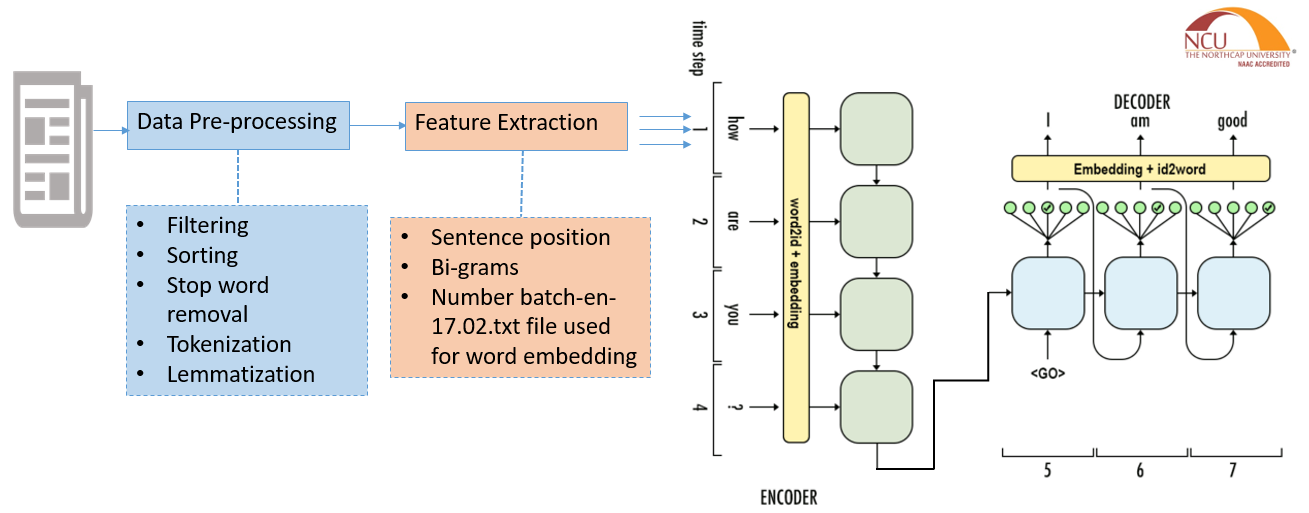
**The summary of background study is presented in table:-**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Method/Approach** | **Evaluation** | **Challenges** | **Features/Accuracy** |
| N.Moratanch & S.Chitrakalat(2017)[1] | Text Summarization using extractive method | 1. Human evaluation 2. Rouge 3. Recall 4. Precision 5. F measure   6. Compression ratio | The main problem in evaluation comes from the impossibility of building a standard against which the results of the systems that have to be compared. | 1)Word level  2)Sentence level |
| Vishal Gupta (2010)[2] | News Document using Extractive Method. | 1) Intrinsic measure(human evaluation)  2) Extrinsic measure(task based performance) | The biggest challenge for text summarization is to summarize content from a number of textual and semi structured sources, including databases and web pages, in the right way (language, format, size, time) for a specific user. | Accuracy of the system is varies from 81% to 92 % |
| Manjula Subramaniam and Vipul Dalal (2015)[3] | Text Summarization using Abstractive Method | Parameters calculated are:-  1) Sentence length  2) Average TF\_ISF (Term FrequencyInverse Sentence Frequency)  3) Sentence position  4) Numerical Data | The generated multiple texts are evaluated and ranked, where the most ranked text is considered. | It uses Rich Semantic Graph techniques |
| N.Moratanch & S.Chitrakalat(2016)[4] | Text Summarization using extractive method | ROUGE methods are:-  1) ROUGE-N  2) ROUGE-L  3) ROUGE-W  4) ROUGESU. | The major issues of abstractive summarization is there is no generalized framework , parsing and alignment of parse trees is difficult. | 1)Single document  2)Multiple document |
| Anitha J., Prof. P. V. G. D. Prasad Reddy and M. S. Prasad Babu (2014)[5] | Text Summarization using Extractive Method | 1)Precission  2)Recall  3)F-score | Hybrid Fuzzy algorithm is more effective that clustering algorithm | It uses fuzzy classifier and Neural Network. (Precision = 0.90, Recall = 0.88) |
| Das and Bandyopadhyay (2010)[6] | Text Summarization using Extractive Method | 1)k-means approach  2)Page Rank standard approach | more than one reference summaries are used for evaluating each system generated summary, but in our work, we have used only one reference summary for summary evaluation. | Precision= 72.15%, Recall=67.32% and FScore= 69.65 % [6] |
| Jagadish S. Kallimani, K. G. Srinivasa and B. E. Reddy (2016)[7] | Text Summarization using Abstractive Method | 1)F score  2)Precision  3)Recall | It uses IE rules and class based templates | F score- 0.815, Precision-0.8642, Recall0.7973, Accuracy- 0.7217 |
| Dr. Latesh Malik (2013)[8] | Single document Summarization using Extraction Method | In sentence similarity feature subject of sentence is matched with subject of title |  | It uses statistical & linguistic feature & also uses Genetic Algorithm |
| Stefan Thomas, Christian Beutenmuller, Xose de la Puente Robert Remus and Stefan Bordag (2015)[9] | Text Summarization using Extractive Method | Rouge Score | Manually inspecting a number of summaries, we notice that very similar sentences recurring often in texts are rarely selected by the Frank algorithm | Sentences are selected using an iterative extension of PageRank calculation on a sentence similarity graph. |
| Renjith S.R, Sony P (2015) [10] | Text Summarization using Extractive Method | Compression ratio | Since an efficient WordNet for Malayalam is notyet implemented,a synset for the corpus under considerationwill be made use of. | It uses TF-IDF techniques |
| Aakash Sinha, Abhishek Yadav and Akshay Gahlot(2018) [11] | Extractive Text Summarization using Neural Networks | ROUGE(Recall-Oriented Understudy for Gisting Evaluation) methods are:-  1)ROUGE-1  2)ROUGE-2 | We have assumed that summary length to be generated should be less than ‘page\_len’. So we will try to improve upon this aspect in the future | For the generation of the summary of a given document, the entire text is broken into pages. The summary length in terms of the number of sentences is fixed and known before summary generation. |
| Shengli Song, Haitao Huang and Tongxiao Ruan [12] | Abstractive text summarization using LSTM-CNN based deep learning | ROUGE (Recall-Oriented Understudy for Gisting Evaluation) | There are many points in ATS models that deserve our attention.  (i) CoreNLP is a very useful tool.  (ii) Determining semantic similarity between phrases is difficult.  (iii) Training deep learning models are a very time-consuming task and determining the threshold δ in the decoder of the model requires manual tuning. | Stanford natural language process tool CoreNLP to pre-process original text |
| Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, Min Sun(2018) [13] | A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss | ROUGE-1, ROUGE-2 and  ROUGE-L F-1 |  | train our extractor and abstracter with 128-dimension word embeddings and set the vocabu-lary size to 50k for both source and target text. |
| Jiacheng Xu, Zhe Gan, Yu Cheng, Jingjing Liu(2019) [14] | Text Summarization using Extractive Method | ROUGE-1, ROUGE-2 and  ROUGE-L F-1 | we will explore better graph encoding  methods, and apply discourse graphs to other tasks  that require long document encoding | 2 models used CNNDM and NYT |
| AR Deshpande, L Lobo(2013)[15] | Text Summarization using Extractive Method | 1)Precission  2)Recall  3)F-score | In future, we would like to improve the system by adding sentence simplification technique for producing summary i.e. it can be used to simplify the sentences which are complex and very large | Maintained a list of maps where each term from document collection is stored in a map with its number of occurrences. A map contains all the synonyms, of the term from the document collection. We have used WordNet dictionary to find synonyms |
| JM Sanchez-Gomez, MA Vega-Rodríguez(2018) [16] | Text Summarization using Extractive Method | 1.ROUGE-N  2.ROUGE-L | As a future research, the approach will be adapted to be applied in Neuro K software 1, which is ane-learning platform based on neurodidactics and social networks principles | Multi-Objective Artificial Bee Colony(MOABC) algorithm |
| S Ryang, T Abekawa(2012)[17] | Text Summarization using Extractive Method | ROUGE score | We also intend to consider efficient features and a score to achieve stable convergence. In addition, we plan to use other methods of function approximation, such as RBF networks | Automatic Summarization using ReinforcementLearning (ASRL) |
| S Dohare, H Karnick, V Gupta(2017)[18] | Text Summarization using abstractive method | 1.Recall  2.Precision  3.F1 scores | The most exciting improve-ments can be done in the summary graph extrac-tion method. Not a lot of work has been doneto extract AMR graphs for summaries. | Proposed method achieves state-of- the-art results compared to the other text summarization routines based on Abstract Meaning Rep-resentation(AMR) |
| Jiamou Sun, Zhenchang Xing, Hao Guo, Deheng Ye, Xiaohong Li, Xiwei Xu, Liming Zhu(2021)[19] | Text Summarization using extractive method | 1.Precision  2.Recall  3.F1 scores  4.ROUGE | Another challenge is that root cause, attack vector and impact may be mentioned several time but not in the exact same expression. | The accuracy of the 384 sampled vendors and attacker types is 80% and 94% respectively |
| [T Shi](https://scholar.google.com/citations?user=muaz7AoAAAAJ&hl=en&oi=sra), [Y Keneshloo](https://scholar.google.com/citations?user=EF0ZoXcAAAAJ&hl=en&oi=sra), [N Ramakrishnan](https://scholar.google.com/citations?user=fMJwG7MAAAAJ&hl=en&oi=sra)(2021)[20] | Text Summarization using abstractive method | 1.ROUGE  2.Bleu | 1.Saliency  2.Fluency  3.Human readability  4.Generate high-quality summaries | 1.Attention-based seq2seq framework  2.Pointer-generator network  3.Intra-temporal attention mechanism  4.Coverage mechanism  5.Intra-decoder attention mechanism  6.Weight-sharing mechanism  7.Beam-search algorithm |

**3.1Dataset-** This dataset consists of reviews of fine foods from amazon. This data includes reviews from time span of Oct 1999 - Oct 2012. There are about 568,454 reviews from 256,059 users describing 74,258 products. In this dataset over 260 users have given more than 50 reviews.

**3.2Methodology-** For building the text summarization model, extractive approach is used. The basic idea of extractive approach is to identify important sections of the text and generate a summary which is a subset of the original text. Initially, we imported the necessary libraries and loaded the data. The next step was to pre-process the data which involved scoring of the reviews and checking for duplicate values. Further we make two classes on the basis of score values. Now in order to work with the text reviews we did text pre-processing by removing stop words and applying word2vec lemmatization. Later, we split the data in train and test set with 80:20 ratio. Now, the main task was to build a LSTM model with embedding layer. For evaluation we used accuracy, F1 score, recall and precision. The working of LSTM model can be seen in figure 1.

For the LSA model, we first imported the necessary libraries and loaded the dataset. We just needed the review column, hence we dropped other columns.Then we used the wordnet lemmatizer followed by tf-idf vectorizer. Then we haved loaded our LSA model and fed our data to it for training. In this we have used first 10 reviews using LSA score. Using this wordclouds were constructed.



**Figure:3.2.1 Long Short Term Memory (LSTM Unit)**

**4.Result-** After applying supervised and unsupervised models, results were computed. The supervised learning is done using LSTM model on the dataset. Unlike standard feed forward neural networks, LSTM has feedback connections and are more controllable networks.

**Table 4.1**

|  |  |  |
| --- | --- | --- |
| S.No. | Method | Value |
|  | F1-Score | 86.8% |
|  | Precision | 80.1% |
|  | Recall | 94.5% |
|  | Accuracy | 79.1% |

The unsupervised learning is applied using LSA and LDA model. Our model was trained using amazon food review dataset in order to visualize our result, word cloud were used. Both LSA and LDA use bag of words as input matrix, this matrix is very sparse and noisy. But in case of LSA Truncated SVD(Singular Value Decomposition) is used as it finds the most valuable information and uses low dimension to represent the same thing.

**5.Conclusion-** Text Summarization is a procedure in which important information is extracted from main text and represented in form of short summary. In this paper, extractive approach of text summarization is used. The objective of this paper was to analyze and compare the supervised, unsupervised learning algorithm in text summarization. This dataset consists of reviews of fine foods from amazon. Reviews include product and user information, ratings, and a plain text review. The idea of extractive approach is to identify important sections of the text and generate a summary which is a subset of the original text.

Refrences

[1] Moratanch, N., and S. Chitrakala. "A survey on abstractive text summarization." *2016 International Conference on Circuit, power and computing technologies (ICCPCT)*. IEEE, 2016.

[2] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." *Journal of emerging technologies in web intelligence* 2.3 (2010): 258-268.

[3] Subramaniam, Manjula, and Vipul Dalal. "Test model for rich semantic graph representation for Hindi text using abstractive method." *International Research Journal of Engineering and Technology* 2.2 (2015): 113-116.

[4] Chitrakala, S., et al. "Concept-Based Extractive Text Summarization Using Graph Modelling and Weighted Iterative Ranking." *International Conference on Emerging Research in Computing, Information, Communication and Applications*. Springer, Singapore, 2016.

[5] Anitha, J., PVGD Prasad Reddy, and MS Prasad Babu. "Error tolerant global search incorporated with deep learning algorithm to automatic Hindi text summarisation." *International Journal of Business Intelligence and Data Mining* 14.3 (2019): 359-380.

[6] Das, Amitava, and Sivaji Bandyopadhyay. "SentiWordNet for Indian languages." *Proceedings of the eighth workshop on Asian language resouces*. 2010.

[7] Kallimani, Jagadish S., K. G. Srinivasa, and B. Eswara Reddy. "Statistical and analytical study of guided abstractive text summarization." *Current Science* (2016): 69-72.

[8] Thaokar, Chetana, and Latesh Malik. "Test model for summarizing hindi text using extraction method." *2013 IEEE Conference on Information & Communication Technologies*. IEEE, 2013.

[9] Thomas, Stefan, et al. "Exb text summarizer." *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2015.

[10] Renjith, S. R., and P. Sony. "An automatic text summarization for Malayalam using sentence extraction." *proceedings of 27th IRF International Conference*. 2015.

[11] Sinha, Aakash, Abhishek Yadav, and Akshay Gahlot. "Extractive text summarization using neural networks." *arXiv preprint arXiv:1802.10137* (2018).

[12] Song, Shengli, Haitao Huang, and Tongxiao Ruan. "Abstractive text summarization using LSTM-CNN based deep learning." *Multimedia Tools and Applications* 78.1 (2019): 857-875.

[13] Hsu, W. T., Lin, C. K., Lee, M. Y., Min, K., Tang, J., & Sun, M. (2018). A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

[14] Xu, Jiacheng, et al. "Discourse-aware neural extractive text summarization." *arXiv preprint arXiv:1910.14142* (2019).

[15] Deshpande, Anjali R., and L. M. R. J. Lobo. "Text summarization using clustering technique." International Journal of Engineering Trends and Technology 4.8 (2013): 3348-3351.

[16] Sanchez-Gomez, Jesus M., Miguel A. Vega-Rodríguez, and Carlos J. Pérez. "Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach." Knowledge-Based Systems 159 (2018): 1-8.

[17] Ryang, Seonggi, and Takeshi Abekawa. "Framework of automatic text summarization using reinforcement learning." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012.

[18] Dohare, Shibhansh, Harish Karnick, and Vivek Gupta. "Text summarization using abstract meaning representation." arXiv preprint arXiv:1706.01678 (2017).

[19] Sun, Jiamou, et al. "Generating Informative CVE Description From ExploitDB Posts by Extractive Summarization." arXiv preprint arXiv:2101.01431 (2021).

[20] Shi, Tian, et al. "Neural abstractive text summarization with sequence-to-sequence models." ACM Transactions on Data Science 2.1 (2021): 1-37.