An Analysis of Text Summarization Approaches using Natural Language Processing

Kavita Namdeo Jadhav[1]

[1] Assistant Professor, Sanmati Engineering College, Washim, (M.S.), India

Abstract: Text Summarization is a Natural Language Processing (NLP) concept to extracts and collects data from the source and summarizes it. There has been a continuous increase in the number of internet users every day, causing a great deal of information that gets stored online. Organizations generating a huge amount of data need data summarization without losing the original meaning of the data. Thus, the process of Text Summarization comes into the picture with its benefits spread over different fields such as Machine Learning, Natural Language Processing, Artificial Learning, Semantics etc. Nowadays, Text summarization is a requirement for many applications as manual summarization of huge amounts of data is difficult, especially with the expanding magnitude of data. Financial research, search engine optimization, media monitoring, question-answering bots, and document analysis all benefit from text summarization. All the e-commerce applications have a review system for the customer, every new user looks at the reviews of previous customers before ordering anything online. Nevertheless, reading long reviews is not easy for everyone. Therefore, there should be a method that can reduce the long reviews to short sentences of limited words depicting the same meaning. Text Summarization can come in handy in this aspect. Many NLP researchers are interested in Text Summarization. This paper extensively addresses the importance of text summarization using NLP and a survey on the various types of text summarization techniques starting from the basic to the advanced techniques. Our aim is to evaluate and convey an abstract view of the present scenario research work for text summarization.

Keywords: Text Summarization, Natural Language Processing, NLP, etc.

I. INTRODUCTION TO TEXT SUMMARIZATION

In the era of e-commerce online shopping has become a common thing as a wide variety of products are available in a single place. The ease of ordering a product and getting the product delivered directly to the user's home at a convenient date and time has attracted many people. Along with these, the discount offers being offered by online shopping sites are making people stick to online ordering. Everyone refers to product reviews before buying a product. Then they can conclude which is the best product to buy among the different products available. Suppose a user needs to buy a laptop. Then he must go through different kinds of laptops available on his budget, make a note of different reviews for each product and choose the best among the available laptops. This is a tedious and time taking task. In addition to this, some users' reviews are so long that the user could get the actual meaning of it only after closely going through the review. Thus, there is a need for minimizing the review to a shorter representative sentence which depicts the same meaning as the whole content. [1]

Organizations generating a huge amount of data needs data summarization without losing the original meaning of the data. Thus, the process of Text Summarization comes into the picture with its benefits spread over different fields such as ML, NLP, Artificial Learning, Semantics etc. Nowadays, the number of internet users continuously increasing every day, causing difficulty in information storage. Most of the generated data is unstructured, so manually extracting meaningful data from it is challenging [2].
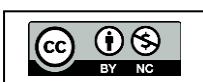
Humans have a constrained ability to comprehend and extract useful information from large amounts of data. It takes a long time for them to grasp the essence of the content. As a result, automatic summarization is a well-known way of addressing such challenges [2]. The objective of text summarization is to gather prominent information from the source by filtering and providing a succinct summary [2]. To date, several techniques for text summarization have been developed. Text summarization techniques can be broadly classified into four categories: input, output, content, and purpose. There are single and multi-document summary options based on the number of documents. Meanwhile, the extractive and abstractive outcomes are based on the summary results. [3]
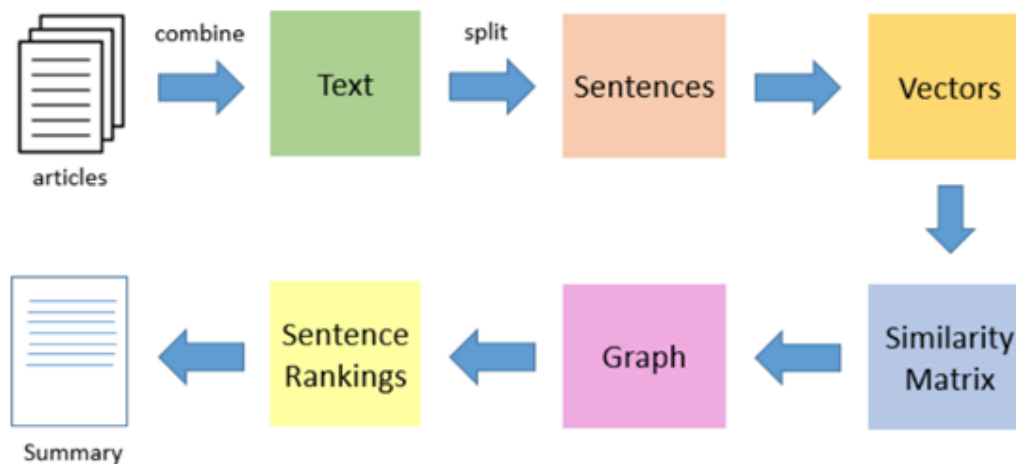

Figure 1: Text Summarization Process

II. TEXT SUMMARIZATION APPROACHES

The Text Summarization Approach is based on the algorithms or programs which reduces the text size and create a summary of our text data. This is called automatic text summarization in machine learning. Text summarization is the process of creating shorter text without removing the semantic structure of text.

There are three approaches to text summarization.

1. Based on the input type
2. Based on the Purpose
3. Based on the output type.

1) BASED ON THE INPUT TYPE:

i. Single Document Text Summarization (SDTS):

In this type of Summarization, the length of the input is short. There will be only a single document given as the input for Summarization. This was used in the early days of Text Summarization.

ii. Multi-Document Text Summarization:

This is a process in which the length of the input on a particular topic is too long and therefore multiple documents are provided as input for a summarization technique. This is often difficult when compared with the SDTS as there is a need to combine the summary of multi documents into a single document. The difficulty here is that there may be diversity in the themes of different documents. An ideal summarization technique often makes condenses the main themes maintaining readability, completeness and without missing the important sentences. [4-6]
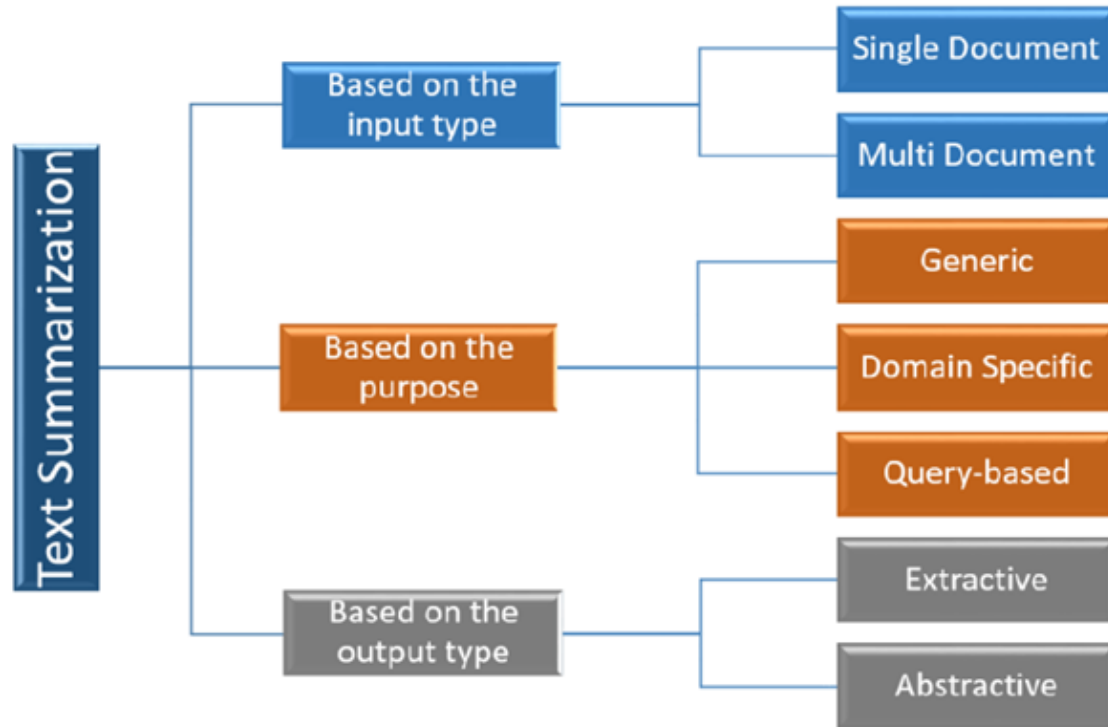
Figure 2: Text Summarization Approaches

2) BASED ON THE PURPOSE:

i.   *Generic Text Summarization:*

The method in which the model makes no inferences about the meaning of the text to be summarized or any knowledge of the domain is called Generic Text Summarization. It makes a generic summary of the whole text, documents, photos, or video clips.

ii. *Domain-Specific Text Summarization:*

In this method of text summarization, the model uses knowledge of a specific domain like scientific documents, and medical documents. This increases the accuracy and thereby gives a more meaningful, concise, and easily understandable summary of the whole text.

iii. *Query-based Text Summarization:*

This method involves taking a query as input and based on that query, the model makes a summary of the text by selecting the sentences and phrases that are very much related to the query given as input. [6 - 8]

3) BASED ON THE OUTPUT TYPE:

i.   *Extractive Text Summarization:*

As the name itself depicts, extractive text summarization is the process in which the sentences are extracted from the whole text which could depict a similar meaning as the whole text but in a more condensed form. Most of the text summarization techniques that are being used nowadays are of an extractive type.

ii. *Abstractive Text Summarization:*

This is a more advanced type of text summarization which involves the formation of phrases or sentences that are not in the text but reflect the same meaning as the complete text. This method is more captivating but at the same time, it is more difficult for the model to form phrases or sentences that could bring the same meaning. [9]

III. TEXT SUMMARIZATION IN NLP

Text summarization is a very useful and important part of Natural Language Processing (NLP). Suppose we have too many lines of text data in any form, such as from articles or magazines or on social media. We have time scarcity so we want only a nutshell report of that text. We can summarize our text in a few lines by removing unimportant text and converting the same text into a smaller semantic text form.

NLP text summarization is the process of breaking down the lengthy text into digestible paragraphs or sentences. This method extracts vital information while also preserving the meaning of the text. This reduces the time required for grasping lengthy pieces such as articles without losing vital information. Text summarization is the process of creating a concise, coherent, and fluent summary of a longer text document, which involves underlining the document's key points.

Text summarization uses AI and NLP to take large amounts of data and extract the most relevant details from the text. The most important points within the document are summarized, while the meaning of the data is maintained. With the volume of data being constantly generated, research analysts do not have the time to read through every word and extract important points manually.

IV. BENEFITS OF NLP TEXT SUMMARIZATION

1. **Scalable and Quick**

Manually summarizing a short document is easy, but what if users have an article or paper that is hundreds or thousands of pages long? Rather than having dozens of employees manually going through thousands of documents, users can automate this analysis with the Accern NoCodeNLP Platform. The text summarization using NLP can analyze all users' input text and source documents and provide a summary text.

2. **Leverages Existing Tools**

To take advantage of text summarization, users do not have to build machine learning infrastructure in-house or hire a data science team. The technology exists and it is accessible to everyone. The summarization of documents and transformation of data, words, and sentences into decisions is possible and already used in a variety of industries with AI / ML / NLP platforms like ours.

The summarization using NLP provides a technology that can effortlessly extract key points from a given document and provide a detailed summary using NLP. The text summarization with NLP algorithms makes text summarization easy to use and available to make the research and business decision-making processes more efficient and actionable.

3. **Understand Your Customers Better**

Users can never go wrong with having a better understanding of customers. The data on customers may come in many forms, such as spreadsheets, chat messages, emails, etc. And of course, it might come in different languages too. In customer relationship management, text summarization refers to compressing
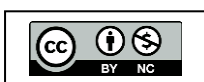
all the above-mentioned customer data, and turning it into abridged summaries that users can present in a meeting or use in other business processes. Since NLP can extract insight from text data, this makes it a perfect tool for keeping track of customer feedback, determining sentiment, whether it is positive or negative, and to what degree. This allows organizations to monitor reviews in real time and flag the most important or time-sensitive comments, provide timely feedback, and ignore irrelevant information.

4. **Summarize the Text in Different Formats**

Whether users are conducting research or launching a new product, one of the first steps is analyzing the market and competitors. Natural language processing helps users to obtain summarized text extracted from competitors' web pages, market research documents, industry-related articles, etc. Having a clear idea of the market and competitors helps users determine actionable steps for presenting products or refining business strategy. This helps users stand out amongst the competitors and maintain a competitive advantage in the market.

The Accern NLP platform can provide users with the most relevant sentences that users can use to communicate products, and important points to focus on and give a deep understanding of the environment. People who are not familiar with extracting data from financial statistics or reports, for example, can use automated text summarization for capturing a synopsis of those statistics and reports.

5. **Ensure all Critical Information is Covered**

The human eye can overlook important details, while standalone software is more accurate. What every reader needs are the ability to extract what is useful to them from any piece of material. The automated text summarizing approach makes it easy for the user to read all the most important sentences in a document.

6. **Run Sentiment Analysis**

Sentiment analysis uses artificial intelligence and natural language processing to identify, extract, and analyze textual data to understand the overall attitude and emotional tone of a text. Once the data is evaluated, a sentiment score will be generated to determine whether the data is positive, negative, or neutral and to what extent. Sentiment analysis is used across many industries, including by hedge funds for analyzing financial news to predict stock market trends and movements.

It is also used by traders and investment bankers to research, extract data on and analyze ESG compliance and mergers and acquisitions news. With sentiment analysis (downloadable white paper), financial teams can even evaluate consumer sentiments around specific companies.

V. NLP TEXT SUMMARIZATION USE CASES

1. **Financial Research with NLP**

Financial and investment decisions require an in-depth investigation and classification of a significant quantity of information. This is true for both individual investors and investment firms. This is where an automatic text summarization designed for evaluating and condensing financial information can come in handy. People who are unfamiliar with extracting data from financial statistics or reports, for example, can use automated text summarization for providing a concise summary of those financial reports. Watch this 3-min video to see how asset managers can leverage the Accern NoCodeNLP Platform to receive fast and accurate insights from text data to fuel timely investment decisions.
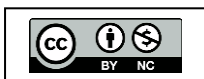
2. **Media Monitoring with NLP**

Assume users need to learn about the current state of an industry from a variety of publications and media. But hardly have time to scan all the headlines, let alone read all of them and get to the meat of their arguments. In this circumstance, text summarization can help users to scan more information by extracting the summary of numerous news articles and other media.

VI. ANALYSIS OF TEXT SUMMARIZATION APPROACHES

In this paper, we have discussed and analysed many prominent works in Text Summarization from the past few years. Earlier works dealt mainly with Single Document Text Summarization. Now recent technologies and high computing power perform a fast, more effective, and more accurate way of processing documents when compared with the earlier methods.

1. **Word Graph Methodology:**

Modi, S., & Oza, R., (2018) discussed Review on Abstractive Text Summarization Techniques (ATST) for single and multi-documents. In this, the technique based on word graphs is separated into two components. The first component is sentencing reduction, followed by sentence combination. This technique involves nodes that represent the information about words and their relation. The word graph technique provides syntactically accurate phrases. The word graph technique creates ungrammatical phrases and is unconcerned with word meaning. [10]

2. **Semantic Graph Reduction Algorithm**

Mridha, M. F., et. al, (2021) discussed A Survey of Automatic Text Summarization: Progress, Process and Challenges. In this, the semantic graph-based approach builds a graph that summarizes the original content by gathering semantic information from words and assigning weights to nodes and edges. This method's strength is producing short, coherent, and grammatically accurate phrases with few networks. This approach is restricted to summarizing material from a single document. [11]

3. **Markov Clustering Algorithm**

Sahoo, D., Bhoi, A., & Balabantaray, R. C., (2018) discussed the Hybrid approach to abstractive summarization. In this construct summary, the Markov Clustering Principle employs a hybrid technique. In this method, sentence ranking is accomplished by combining linguistic norms with the best-fitting sentences inside a cluster to construct new sentences. Sentences are grouped using semantic and statistical variables in the Markov Clustering Principle to produce highly linked sentences. The accuracy of the summary provided by this depends upon the quality of the sentence compression technique. [12]

4. **Encoder-Decoder Model**

Shinde, M., (2021) discussed the Techniques and Research in Text Summarization. In this, the encoder converts the input sentence sequence into a context vector, and the decoder converts the processed input into comprehensible output. The encoder-decoder strength is that it addresses the vanishing gradient issue. The approach requires an extensive dataset that takes a long time to train. [13-14]

5. **Pegasus**

Gupta, A., Chugh, D., & Katarya, R. (2021) discussed automated News Summarization Using Transformers. In this approach, significant lines are eliminated from the input text and compiled as separate outputs. The
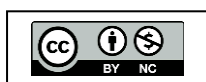
strength of this method is that it selects phrases based on relevance rather than randomness. Pegasus may need post-processing to remove errors and enhance summary text output. [15-17]

6. **Summarization with Pointer Generator Networks**

   Anh, D. T., & Trang, N. T. T. (2019), discussed Abstractive text summarization using pointer-generator networks with pre-trained word embedding. This method employs a hybrid approach, producing words from a predefined vocabulary and replicating words by pointing. This method focuses on resolving the issue of out-of-vocabulary terms. The essence of this technique is to present a summary based on the source content, rather than adding new terminology. [18-19]

7. **Genetic Semantic Graph-based Approach**

   The approach generates a semantic graph from the source text, with graph nodes representing predicate-argument structures (PASs) and graph edges representing semantic similarity weights. The merit of this method is that it reduces redundant information by combining comparable information across documents. The shortcoming of this technique is that it fails to recognize redundant phrases that are semantically similar, resulting in an inadequate final summary. [20]

8. **TF-IDF Approach**

   TF-IDF algorithm calculates the frequency of words in documents and generates metric values. Finally, phrases with a higher metric value are included in the result. The TF-IDF algorithm is quick to compute and has an excellent ability to determine the relevance of phrases. The main disadvantage of TF-IDF is that lengthier sentences get a higher metric score due to the terms' higher occurrence in the sentences. [21]

9. **Approach based on Clustering**

   The clustering technique focuses on grouping texts and creating cluster-level summaries. The clusters are generated using word weight, sentence location, phrase length, sentence centrality, and proper nouns. The significance of clustering resides in its ability to exclude redundant phrases from summaries automatically. The drawback of the clustering approach is that the summarized phrases are not synchronized, and comparing the similarity between clusters is a challenging operation. [22]

10. **Neural Network Approach**

    This method works by first training the neural network, and then the trained network selects the essential phrases that should be included in the summary in the same manner that a person would. The fundamental advantage of neural networks is their ability to change characteristics based on the needs of the user. It takes an excessive amount of time to train a neural network. [25-26]

11. **Approach based on Machine Learning**

    The machine learning technique is classified into two types: supervised, in which documents and summaries are supplied, and unsupervised, in which just documents are provided, and the machine learns by evaluating them. The benefit of the Machine Learning technique is that it is simple to construct and train the model. The limitation is that significant terms often occur in the test dataset but not in the training dataset and are ignored. [27 - 29]
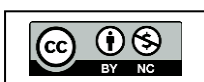
VII. CONCLUSION

Text summarization is a branch of NLP that focuses on shortening texts and making them more readable for users. With an excess of data accessible on the internet and the necessity to comprehend it in order to save the reader's time, text summary techniques are utilized. This paper provides a quick overview of text pre-processing, used to clean data to do effective summarization. Then it summarizes the many types of text summarizing approaches, categorizing them according to input, output, content, and purpose. The paper's primary emphasis is on extractive and abstractive text summarizing algorithms based on output. Extractive summarization summarizes by simply extracting information from the input text.

Abstractive summarization is a more complicated method because it summarizes the text in its language. The abstractive technique produces better and more semantically connected summaries. Readers would benefit significantly from an overview of the benefits and drawbacks of different techniques, as well as a concise explanation. Text summarization techniques can be applied helpfully depending on the user's needs.

REFERENCES

[1] Ravali Boorugu, Dr. Gajula Ramesh, and Dr. Karanam Madhavi, "Summarizing Product Reviews Using NLP Based Text Summarization", International Journal of Scientific & Technology Research, Volume 8, Issue 10, October 2019. ISSN 2277-8616.

[2] Ravali Boorugu and Dr. G. Ramesh, "A Survey on NLP based Text Summarization for Summarizing Product Reviews", Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.

[3] Manasaveerashyva Y N, Prathibha B S, "A Survey on Text Summarization Using Natural Language Processing", IJIRT, Volume 9, Issue 1, ISSN: 2349-6002, June 2022.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean," Distributed Representations of Words and Phrases and their Compositionality," arXiv:1310.4546v1 [cs.CL], 2013.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781v3 [cs.CL], 2013.

[6] Shi Ziyan "The Design and Implementation of Domain-specific Text Summarization System based on Co-reference Resolution Algorithm hm" 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery.

[7] Paul Gigioli, Nikhita Sagar, Anand Rao, Joseph Voyles "Domain-Aware Abstractive Text Summarization for Medical Document s" published during 2018 IEEE BIBM.

[8] Niladri Chatterjee, Amol Mittal and Shubham Goyal's "Single Document Extractive Text Summarization Using Genetic Algorithms" (2012)

[9] Amol Tandel, Brijesh Modi, Priyasha Gupta, Shreya Wagle and Sujata Khedkar's "Multi-document text summarization - A survey" 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).

[10] Talukder, M. A. I., Abujar, S., Masum, A. K. M., Akter, S., & Hossain, S. A. (2020, July). Comparative Study on Abstractive Text Summarization. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE

[11] Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., & Kabir, M. M. (2021). A Survey of Automatic Text Summarization: Progress, Processand Challenges. IEEEAccess, 9, 156043-156070

[12] Sahoo, D., Bhoi, A., & Balabantaray, R. C. (2018). Hybrid approach to abstractive summarization. Procediacomputer science, 132, 1228-1237.

[13] Shinde, M., Mhatre, D., & Marwal, G. (2021, March). Techniques and Research in Text Summarization-A Survey. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 260-263). IEEE.

[14] Wikipedia Contributors. (2020). Multi- DocumentSummarization— Wikipedia, the Free Encyclopedia. Accessed: Oct. 8, 2021. https://en.wikipedia.org/w/index.php?title=Multi- document_ summarization%&oldid=986613170
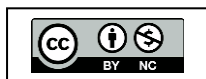
[15] Gupta, A., Chugh, D., & Katarya, R. (2021). Automated News Summarization Using Transformers. arXiv preprint arXiv:2108.01064

[16] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328- 11339). PMLR.

[17] Oliveira, L. M. R., Busson, A. J. G., Carlos de Salles, S. N., dos Santos, G. N., & Colcher, S. (2021, November). Automatic Generation of Learning Objects Using Text Summarizer Based on Deep Learning Models. In Anais do XXXII Simpósio Brasileiro de Informática na Educação (pp. 728- 736). SBC.

[18] Anh, D. T., & Trang, N. T. T. (2019, December). Abstractive text summarization using pointer- generator networks with pre-trained word embedding. In Proceedings of the tenth international symposium on information and communication technology (pp. 473-478).

[19] Boutkan, F., Ranzijn, J., Rau, D., & van der Wel, E. (2019). Point-less: More abstractive summarization with pointer-generator networks. arXiv preprint arXiv:1905.01975.

[20] Khan, A., Salim, N., & Kumar, Y. J. (2015, October). Genetic semantic graph approach for multi- document abstractive summarization. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC) (pp. 173- 181). IEEE.

[21] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268.

[22] Deshpande, A. R., & Lobo, L. M. R. J. (2013). Text summarization using clustering technique. International Journal of Engineering Trends and Technology, 4(8), 3348-3351.

[23] Jewani, K., Damankar, O., Janyani, N., Mhatre, D., & Gangwani, S. (2021, March). A Brief Study on Approaches for Extractive Summarization. In 2021 International Conference on Artificial Intelligence andSmart Systems (ICAIS) (pp. 601-608). IEEE.

[24] Akter, S., Asa, A. S., Uddin, M. P., Hossain, M. D., Roy, S. K., & Afjal, M. I. (2017, February). An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. In 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 1-6). IEEE.

[25] Moratanch, N., & Chitrakala, S. (2017, January). A survey on extractive text summarization. In 2017 international conference on computer, communication, and signal processing (ICCCSP) (pp. 1-6). IEEE.

[26] Andhale, N., & Bewoor, L. A. (2016, August). An overviewof text summarization technique. In 2016 International Conference on Computing Communication Control and automation (ICCUBEA) (pp. 1-7). IEEE.

[27] Kumar, A. K. S. H. I., & Sharma, A. D. I. T. I. (2019). Systematic literature review of fuzzy logic-based text summarization. Iranian journal of fuzzy systems, 16(5),45-59.

[28] Patel, R., Thakkar, A., Makwana, K., & Patel, J. (2017, March). Comprehensive and Evolution Study Focusing on Comparative Analysis of Automatic Text Summarization. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 383-389). Springer, Cham.

[29] Lagrini, S., Redjimi, M., & Azizi, N. (2017). Automatic arabic text summarization approaches. International Journal of Computer Applications, 164(5), 31-37.