2024 3rd International Conference for Advancement in Technology (ICONAT)
Goa, India. Sep 13-14, 2024

Automatic Text Summarization using Text Rank Algorithm

Rakhi Dumne
Department of Comp. Science and Engineering
Walchand College of Engineering
Sangli, India
rsdumne28@gmail.com

Nitin L. Gavankar
Department of Comp. Science and Engineering
Walchand College of Engineering
Sangli, India
nitin.gavankar@walchandsangli.ac.in

Madhav M. Bokare
Department of Computer Institute of Technology and Management
Nanded, India
bokaremadhav@gmail.com

Vivek N. Waghmare
Department of Computer Engineering
PVG's College of Engineering & S.S. Dhamankar Institute of Management
Nashik
dr.vnwaghmare@gmail.com

2024 3rd International Conference for Advancement in Technology (ICONAT) | 979-8-3503-5417-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICONAT61936.2024.10775241

***Abstract*—Automatic text summarization has been emerged as a valuable tool for quickly locating the significant information in vast text with minimal effort. The practice of constructing a summarized form of a text document that retains significant information and also withstand the overall meaning of the source text is known as text summarization. This study mainly concentrates on the extractive text summarization technique wherein the text summarization is carried out on single document as well as on multi document text. Further, the application is extended by implementing document summarization and URL summarization. Here, the sentence extraction method from the input text forms the basis of proposed text summarization technique. During sentence extraction, the weights are assigned to sentences which act as rank of these sentences, using page rank algorithm. In this study, extraction technique has been implemented to extract sentences having higher rank from the given input text. The study mainly focuses on obtaining high rank sentences from the given document in order to generate a high-quality summary of the input text.**

Keywords— Extractive summarization, natural language processing (NLP), Text Rank algorithm, text summarization

I. INTRODUCTION

In present scenario, a large amount of data is being generated on the internet every day. As a result, a better mechanism is required for extracting important information quickly and effectively which has been a major challenge. Text summarization is one of the strategies for finding the most significant and meaningful information in a document or set of linked documents and summarizing it into a shorter version while preserving the overall meaning [1][6]. Text summarizing reduces the amount of time needed to read a lengthy document and takes care of the space problems that are usually encountered with big amounts of data. Extractive text summarization and abstractive text summarization are the two main categories of text summarizing techniques [1]. Automatic text summarizing using extractive text summarization technique is a challenge that is typically broken down into two sub problems: single document and multi document text summarization. Typically, one document is used as the input for single document text summarizing. Further, by applying suitable technique a summary data is generated from the given input while holding its overall meaning. However, in case of multi document text summarization, multiple documents related to the similar subject are given as input and related summary is generated [9].

Various algorithms have been discussed in the literature for text summarization. Some of the challenges in text summarization are word embedding, extraction of key sentences etc. In order to address these challenges, in this study, the Text Rank Algorithm has been proposed for automatic text summarization which is an extractive text summarization approach. The Text Rank Algorithm uses graph based approach for sentence ranking. Here, after preprocessing, each sentence is converted into a vertex as shown in Fig. 1 where the vertices 1-7 represents the sentences and the edges, a line connecting two vertices represents the similarity weight of these sentences [2].

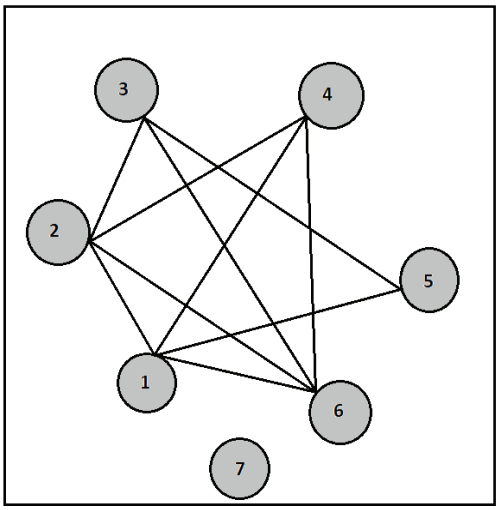

Fig. 1. Sample graph representing similarity [8]

II. LITERATURE SURVEY

The text summarizing techniques used in earlier studies are described in this section. One of the most significant areas of application for natural language processing is text summarization. There are two categories of text summarization techniques: Summary that is both abstract and extractive [1][3]. Extraction of text summarization is the process of removing pertinent sentences from the supplied

text. In order to extract the important content from the input text during extractive text summarization, linguistic and statistical features of paragraphs are typically exploited [3][4]. Whereas, in the abstractive text summarization it comprehends the document's primary notion and its meaning [4]. Further, it interprets the text to find the new notion in the document using the linguistic method. The resultant output will be the most shortened version of the input document [3].

In few studies, summarization of scientific texts has been done based on the important attributes such as, phrase frequency, key words, important phrases, and location of the text [1]. Most of the earlier studies focuses on generating a summary of a single input document. Further, in few studies extractive text summarization has been discussed, in which essential sentences have been extracted by measuring words and phrases frequency that provides a useful indicator of their significance [2]. In 1958, Baxendale began his study on extractive summarization at IBM. Using the text's location, he has extracted crucial sentences [1].

However, Accuracy verification has been a major challenge in the text summarization. Verifying accuracy of the generated summary is calculated manually or human generated summary which is highly subjective. In the proposed system, accuracy verification has been done by using the online software which uses the cosine similarity method in order to calculate the similarity values between the input text.

In addition, many researchers have worked on the topic text summarization using various other algorithm such as Term Frequency Algorithm, Sum Basic Algorithm, Page rank algorithm, Latent semantic indexing etc. which falls under the unsupervised learning algorithm [13].

A. Term Frequency Algorithm

The abbreviation TF-IDF stands for Term Frequency — Inverse Document Frequency. This technique can be used to figure out how many words are present in a group of texts. Here, each word is given a score to represent its weight within the corpus and document. This method is commonly used in text mining and information retrieval [13]. This algorithm has an advantage of easy computation for generating the summary. However, TF-IDF is built on the bag-of-words paradigm, it is unable to account for semantics, co-occurrences across many documents, etc. Whereas, in the present study algorithm used considers the semantic as well as syntactic meaning of a words during word embedding.

B. Sum Basic Algorithm

Sum Basic algorithm is a multi-document text summarization algorithm [10]. It does not support the single document text summarization. In the proposed study, single as well as multi documented text summarization has been implemented in one application itself. In sum basic, the idea is to use more commonly appearing terms in a document than less frequently occurring words in order to produce a summary that is more likely to appear in human abstractions. It generates n-sentence summaries, where n is the number of phrases the user specifies [10].

C. Page Rank Algorithm

Websites are ranked in search engine results using the PageRank (PR) algorithm developed by Google. One of Google's original founders, Larry Page, is accredited for the creation of PageRank [15]. The importance of website pages is measured by using PageRank. Google uses other algorithms as well, but this is its initial and most well-known algorithm. When someone clicks on a random link, the PageRank algorithm creates a probability distribution that shows how likely it is that they will get on a particular page.

D. Latent Segmantic Indexing

The relationship between a group of documents and the terms contained are analyzed using latent sematic indexing. This uses a mathematical concept known as Singular Value Decomposition (SVD) to compute a collection of matrices that represent document similarity. The mathematical approach of the Singular Value Decomposition is used in Latent Semantic Analysis to detect patterns of links between phrases and concepts. This is formed by the idea that words those appear in similar settings have comparable meanings [16].

Depending upon input type, output type and based on their purpose, text summarization has been classified into different types as shown in Fig. 2.

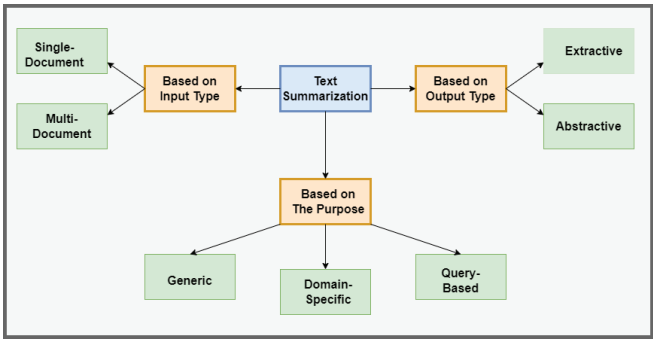

Fig. 2. Types of text summarization [17]

Here in this study the proposed methodology uses the Text Rank Algorithm for text summarization which also falls under the unsupervised learning method. In most of the earlier studies, Single paragraph text summarization has been implemented whereas in the present study single as well as multi paragraph text summarization has been implemented effectively and an easy interface has been provided to the user. Due to its graph based approach the Text Rank Algorithm produces a good outcome in terms of text summarization and which is also language independent [2][4]. In traditional studies, Word2Vec method has been used for the word embedding's which relies on local information of words in the text summarization process. However, the Text Rank Algorithm uses pre trained word embedding dataset which is trained on large dataset. To obtain word vectors, it incorporates global statistics (word co-occurrence) in addition to local information [3]. It also considers the semantic as well as syntactic meaning of a words during word embedding. Considering all these important features of Text Rank Algorithm, the proposed methodology uses Text Rank Algorithm for text summarization in order to improve the performance of the proposed system.

III. PROPOSED METHODOLOGY

Text Rank Algorithm is an extractive and unsupervised text summarization technique**.** It is a graph-based text ranking model for determining the most relevant phrase and keywords in an input text which helps to improve the

accuracy. Overall methodology of the proposed system is shown in the fig 3:

Authentication is the process of verifying the identity of a user. Here, user can see all the available services after login to the system. In order to restrict unauthenticated access every user must register to the system.

Proposed methodology works in five different modules:

- Paragraph Summarizer
- Document Summarizer
- URL Summarizer
- Web Scrapper
- Word Dictionary

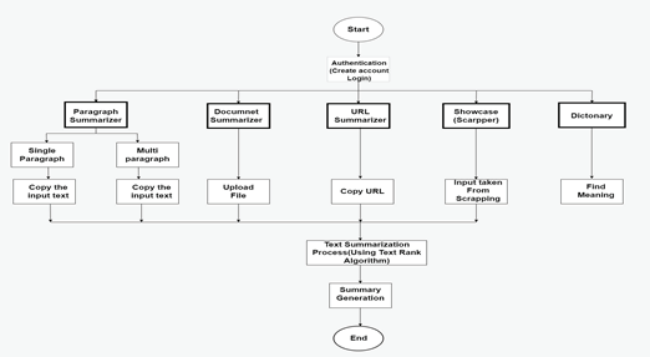

Fig. 3. Proposed Methodology for Text summarization

A. Paragraph Summarizer

In paragraph summarization, user can generate the summary of a single as well as multi paragraph text. This will help the user to read the summary and grasp the best knowledge out of it simultaneously. In most of the earlier studies, only single paragraph summarization interface was available to the user whereas in the present study a multi user paragraph summary interface has been provided which facilitates multi paragraph summarization.

B. Document Summarizer

In Document summarizer, generally the format of the document is restricted to any one of the format. In this study, user can directly upload the files in the different format such as .docx, .pdf, .txt. However, the user can specify 'n' number of sentences required in the summary. The summary is generated by using the Text Rank algorithm and top 'n' high ranked sentences are displayed in the summary. Furthermore, it provides the facility of email where user can send the summary via mail. This email feature is also applicable for the URL summarizer which has been a challenge for the backup of the summary.

C. URL Summarizer

In URL summarizer module, it provides the facility of directly copying the desired URL to the given text box, information will be automatically fetched from the URL and the 'n' number of sentences summary will be displayed. Here, Sumy python library has been used for extracting summary from HTML pages or plain texts. Sumy library helps to avoid the images, videos etc. available on the web pages and extract only the text present on the respective page.

D. Web Scraper

Web Scrapper is the module wherein the summary is generated by using the web scrapping tool. Web scraping is a mean of extracting information in the form of text from the given URL automatically. In showcase model, after the summary of all six technologies is generated, Multi-documented text summarization has been implemented by giving the input text of all 6 technologies as discussed above, then all the data will be appended one after the other and further it serves as the input text to the Text Rank Algorithm.

E. Word Dictionary

Defining of words plays a very important role in the text summarization. As summary is shortened as compared to its original text, it is very important to know the meaning of each and every occurrence of the word in summary. Here, dictionary feature is provided in the application itself for better understanding of the summary and also to get the best from it.

The Text Rank algorithm receives the input text. The sentences in the complete text are mostly ranked by the Text Rank algorithm. Figure 4 illustrates the overall text summarizing procedure using the Text Rank Algorithm:

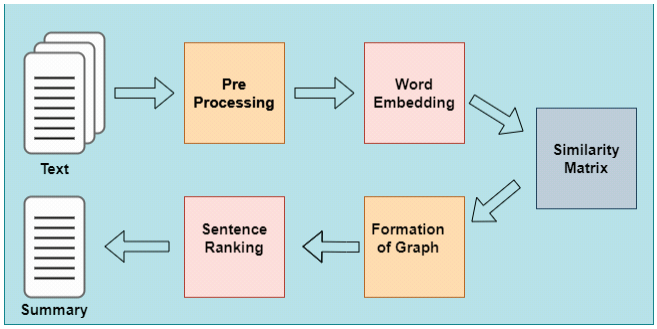

Fig. 4. Text summarization process

- **Preprocessing:** This is the first step in the text summarization process. Cleaning of data from the document is necessary. Since input data consists of vast information along with useful or required information [12]. Further, it is also important to format the data properly in order to achieve better outcomes from the proposed technique. Here in this study the preprocessing procedure carried out is shown in Fig 4.

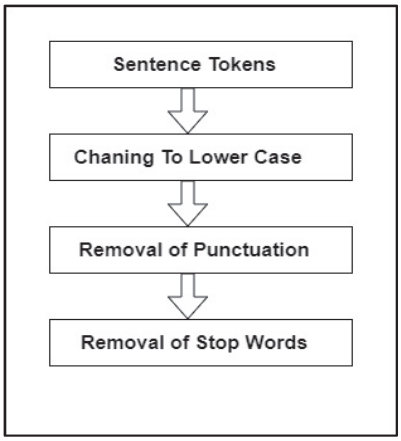

Fig. 5. Pre Processing Flow Diagram [3]

- **Word Embedding's:** The representation of words is usually in the form of a real-valued vector, which is referred to as word embedding. The vectors for the input sentences are calculated by using word embedding technique. The word embedding's most significant use is in encoding a word's meaning and predicting words that are next to each other in the vector space and may have similar meanings [3].

In this study, GloVe data set has been used for vector representation of words available at Kaggle [11]. The use of global statistics to obtain the word vectors is a key component of the dataset. Individual words are stored in a vector space as a real-valued vectors.

- **Similarity Matrix:** Similarity matrix is formed based on a vectors, in which we calculate the similarities between the sentences, then apply the cosine similarity approach. In order to determine how similar, the two sentences are, Cosine Similarity has been used where in values of matrix are filled using cosine similarity scores [3] [12].

  Here, the similarity between the 2 sentence is calculated using following relation:

$$Cos\ (x, y) = x.\ y\ /\ ||x||\ *\ ||y|| \quad (1)$$

- **Formation of Graph:** Once the similarity matrix is derived, further it is converted into graph. A graph is defined as a collection of nodes i.e. vertices and identifiable pairs of nodes which is known as edges or link. The nodes of this graph represent the sentences, and its edges show how similar the sentences are to one another [3].

- **Sentence Ranking:** In order to rank the sentences in the graph, the graph is provided as an input to the Page Rank Algorithm. Although it is not the only algorithm employed by Google to arrange search engine results, it was the first and is the most well-known [18]. It is employed to figure out a web page's weight. In order to choose the appropriate phrases from a list of tokenized sentences, the scores are listed in descending order with their key value used as an index.

IV. RESULT AND DISCUSSION

In the field of text summarization, verifying accuracy has been a major challenge. The techniques such as the comparison between the size of the input text and the output text summary, manual check or expert generated summary which is very subjective in nature are studied earlier. In the proposed text summarization, the model is tested on the various documents having different contents. All these text are different in size. The number of lines in input document varies from 20 to 100 lines in order to test the accuracy of the proposed module on small as well as large input document.

The obtained text summary is compared with the summary generated by standard tool named as SMMRY. SMMRY is an online tool which is used to summarize articles and text. Using an online software program which make use of cosine similarity to determine the degree of similarity between the two documents, the similarity between the proposed system generated summary and the SMMRY generated summary is determined after the summary has been generated. The most effective similarity measurement technique for text summary is cosine similarity. Similarity between the system generated and SMMRY generated is shown in the table 1. Further, for better visualization a graph is plotted considering number of input text against the similarity percentage, Fig. 6.

Once the average similarity of the 50 sample input text has been calculated, the next task is to calculated the average of these sample document. This step helps to derive the approximate accuracy of the proposed system. The results show the accuracy of the proposed technique is around 90.00%.

Based on the ranking of these sentences, the present model produces a summary of n number of sentences. However, the user may choose n, the flexible number of sentences in the generated summary, taking into account the significance and length of the input content. Here, extractive text summarization is implemented using Python 3.7 and NLTK.

TABLE I. AVEVERAGE SIMILARITY

Input Text	*Similarity (Percentage)*	*Input Text*	*Similarity (Percentage)*
1	95.1	26	86.6
2	86.7	27	90.2
3	84.3	28	91.7
4	95.8	29	90.0
5	92.9	30	88.9
6	91.8	31	92.6
7	84.3	32	94.7
8	93.0	33	86.5
9	92.5	34	92.8
10	89.4	35	96.9
11	87.1	36	86.7
12	88.2	37	91.7
13	88.3	38	90.4
14	93.5	39	84.1
15	89.0	40	87.3
16	86.1	41	92.4
17	89.6	42	88.1
18	96.7	43	84.9
19	89.4	44	95.5
20	85.4	45	89.2
21	90.0	46	90.8
22	89.0	47	86.3
23	87.1	48	90.1
24	89.3	49	96.0
25	91.2	50	92.6

Further, User Interface (UI) is an important aspect which facilitate the user. Simple and effective UI encourages the user to use the proposed system. A snapshot of UI is shown in Fig. 7.

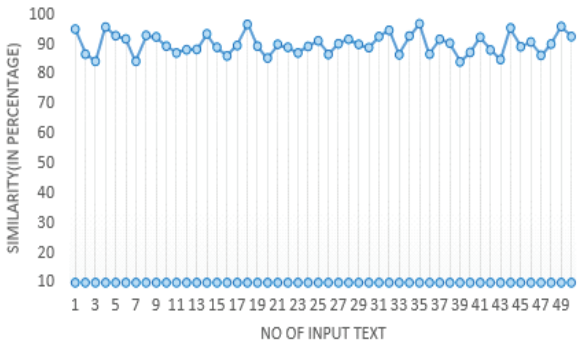

Fig. 6. Line graph for representing similarity values

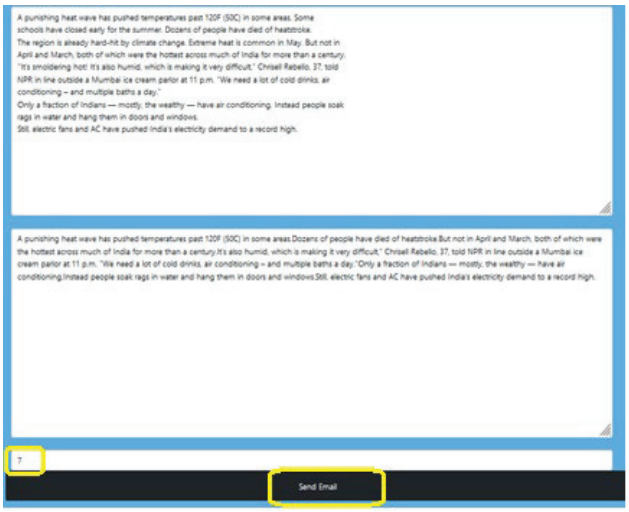

Fig. 7. Snapshot UI- Document Summarization

The proposed system is further integrated to generate summary from 1) given URL – URL summarization and 2) Web Scrapper, which is multi documented text summarization module.

1. **URL Summarization:** In URL summarizer module, it provides the facility of directly copying the desired URL to the given text box, information will be automatically fetched from the URL and the 'n' number of sentences summary will be displayed. Here, Sumy python library has been used for extracting summary from HTML pages or plain texts. Sumy library helps to avoid the images, videos etc. available on the web pages and extract only the text. Here in this module user copy the url with n numbers of sentences required in the generated summary, the summary has been generated using proposed text summarization model.

2. **Web Scrapper:** Web Scraper is the module wherein the summary is generated by using the web scrapping Web Scraper model provides the user interface where web scrapping is used for information extraction from the web pages and respective extracted information is given to the proposed text summarization model algorithm as an input data and then summary has been generated.

The proposed text summarization algorithm is also found effective for URL summarization and using Web Scrapping multi document summarization.

V. CONCLUSION AND FUTURE SCOPE

The main objective of this study is to develop a technique for obtaining the summary of the input text which is useful in various applications such as understanding very long technical and non-technical articles. Current research is focusing on improving accuracy of the summarized text using modern tools and techniques. Here, by using the automated tool, the similarity between the documents has been calculated. The proposed text summarization technique using Text Rank algorithm is used for generating summary of any input text which includes different five models such as paragraph, document, URL and scrapper summarization. All these five services are available to the user in one application providing the better user interface. The experimental results show that the accuracy of the proposed technique is around 90.00%.

In future the accuracy of the proposed system may be tested by considering different articles such as news, blogs on social media, etc. The accuracy of the technique may further be improved by integrating the machine learning approache.

REFERENCES

[1] Madhuri, J.N. and Kumar, R.G., 2019, March. Extractive text summarization using sentence ranking. In *2019 International Conference on Data Science and Communication (IconDSC)* (pp. 1-3). IEEE.

[2] Rahimi, S.R., Mozhdehi, A.T. and Abdolahi, M., 2017, December. An overview on extractive text summarization. In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)* (pp. 0054-0062). IEEE.

[3] Tanwi1, Satanik Ghosh1, Viplav Kumar1, Yashika S Jain1, Mr. Avinash 2019, April. Automatic Text Summarization using Text Rank, In 2019 International Research Journal of Engineering and Technology (IRJET)

[4] Gunawan, D., Harahap, S.H. and Rahmat, R.F., 2019, November. Multi-document summarization by using textrank and maximal marginal relevance for text in bahasa indonesia. In *2019 International Conference on ICT for Smart Society (ICISS)* (Vol. 7, pp. 1-5). IEEE.

[5] Ashna Jain,2019, April.Automatic Extractive Text Summarization using TF-IDF.

[6] Janjanam, P. and Reddy, C.P., 2019, February. Text summarization: an essential study. In *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)* (pp. 1-6). IEEE.

[7] Zaware, S., Patadiya, D., Gaikwad, A., Gulhane, S. and Thakare, A., 2021, June. Text summarization using tf-idf and textrank algorithm. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1399-1407). IEEE.

[8] Rahimi, S.R., Mozhdehi, A.T. and Abdolahi, M., 2017, December. An overview on extractive text summarization. In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)* (pp. 0054-0062). IEEE.

[9] Fakhrezi, M.F., Bijaksana, M.A. and Huda, A.F., 2021. Implementation of Automatic Text Summarization with Text Rank Method in the Development of Al-Qur'an Vocabulary Encyclopedia. *Procedia Computer Science*, *179*, pp.391-398.

[10] Kulkarni, A.R. and Apte, M.S., 2002. An automatic text summarization using feature terms for relevance measure. *IOSR J. Comput. Eng*, *9*, pp.62-66.

[11] Saggion, H. and Poibeau, T., 2013. Automatic text summarization: Past, present and future. In *Multi-source, multilingual information extraction and summarization* (pp. 3-21), Springer, Berlin, Heidelberg.

[12] Dr. Annapurna P Patil, Shivam Dalmia, Syed Abu Ayub Ansari, Tanay Aul, Varun Bhatnagar, "Automatic Text Summarizer ", International Conference on Advances in Computing, Communications and Informatics, IEEE (2014).

[13] Jayashree R, Shreekantha Murthy K, "Categorized Text Document Summarization in the Kannada Language by Sentence Ranking" ,

12th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE (2012).

[14] Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar, "Automatic Text Summarization of News Articles ", International Conference on Big Data, IoT and Data Science (BID) Vishwakarma Institute of Technology, Pune, Dec 20-22 IEEE (2017).

[15] Prachi Shah, Nikitha P. Desai, "A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages ", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (2016).

[16] D. Gunawan and A. Amalia, "Review of the recent research on automatic text sum-27 marization in bahasa indonesia," in 2018 Third International Conference on Informatics and Computing (ICIC), Oct 2018, pp. 1–6.

[17] Min-Yuh Day, Chao Yu Chen, "Artificial Intelligence for Automatic Text Summarization", International Conference on Information Reuse and Integration for Data Science IEEE (2018).

[18] Prachi Shah, Nikhita P Desai, "A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (2016).

[19] Online resource, Multi-Document summarization Available at: https://paperswithcode.com/task/multi-document-summarization

[20] Online resource, Dataset: glove.6B.100d downloaded from kaggle. Available at: https://www.kaggle.com/rtatman/glove-global-vectors-for-word-representation

[21] Online resource, Page Rank Algorithm Available at: https://www.geeksforgeeks.org/page-rank-algorithm-implementation/

[22] Online resource, latent sematic indexing Available at: https://en.wikipedia.org/wiki/Latent_semantic_analysis

[23] Online resource, Page Rank Algorithm Available at: https://devopedia.org/text-summarization

[24] Srinath, K.R., 2017. Page Ranking Algorithms–A Comparison. *International Research Journal of Engineering and Technology (IRJET)*.