2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-6654-9400-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAECT57570.2023.10117938

A Novel Approach to Text Summarization of Document using BERT Embedding

Kumkum S. Adwani
Department of Computer Science & Engineering,
Govt. College of Engineering,
Amravati, India
Email: kumkum29699@gmail.com
(Corresponding author)

Prof. P. P. Shelke
Department of Computer Science & Engineering,
Govt. College of Engineering,
Amravati, India
Email: shelke.prajakta@gcoea.ac.in

***Abstract*— This study examines and compares existing studies on various text summaries of documents and the processes associated with them. Despite the fact that the literature contains a large number of research contributions, we have critically and fully analysed current research and review papers that are relevant to text summary of document systems. The various approaches are classified based on the main ideas used in their procedures. The emphasis is on the concept utilised by the concerned authors, the approach used for experiments, and the performance evaluation measures. The researchers' claims are also emphasised. Our findings from the exhaustive literature review are presented together with the detected flaws. This study is very important for the comparative examination of various text summarizer approaches, which is necessary for addressing associated difficulties.**

On a review of the literature, we developed our own way in which we constructed a programme in Python and wrote it in the Spyder console using the Anaconda culture. The following standard libraries are used: NLTK for text processing, TensorFlow hub for getting the BERT pretrained model, correlation from the Sklearn library, and certain other standard libraries such as Panda for importing the dataset, Numpy, and so on. The data set must be imported after all of the libraries have been included. To summarise, the Kaggle datasets news summary training dataset was utilised. There are headlines, the full text, a synopsis, and links to related news stories for each story in the collection. In this experiment, the first 100 articles from the news summary collection were used. We analysed the outcome using the Rouge scoring system on both the created summary and the original summary. Rouge's average score for the first 100, 50, and 30 documents. The results obtained are encouraging.

Keywords*— *Artificial Intelligence, Text Summarization, Machine Learning, Page Rank, Rogue

I. INTRODUCTION

Because of the growth of social media, online education services, and all professional fields, the amount of knowledge accessible over the internet has expanded in recent years. The main format in which this information is growing is textual data. As a result, the key problem now is processing and understanding such a large volume of data. Condensing the textual content, also known as summarising [2], which is the act of minimising the amount of textual data, is the only option to address this problem. Condensing information into brief summaries is challenging, though. It necessitates having in-depth understanding of the content being summarised. This problem can be resolved using one NLP technique, text summarization, which is discussed further in this section [1].

The overview of the current methodologies is the primary topic of this study. Although several review articles for the text summarising of documents utilising BERT embedding systems have been published in the literature, we have critically examined and methodically described the most recent, significant, and important works. The potential remedy for text summarization systems is also provided in this research. This document's layout is as follows. Section 2 has a systematic presentation of the literature review, which includes a summary of the work experience in addition to key characteristics of the linked techniques. Section 3 sets the stage for the conceptualization of the issue and the discovered problem. Section 4 describes in detail how the selected strategy works. Section 5 reports the experimental findings, while Section 6 provides the perspective.

II. SURVEY OF EXISTING APPROACHES

The available research that is pertinent to text summarizer systems and the mechanics underlying them is critically analysed in this part. Although there are many research contributions in the literature, we have only examined the most recent, significant, and useful study and review publications here. The current methodologies are classified into many categories based on the underlying ideas used in the procedures. The authors' concepts, the platform they employed for their experiments, and the effectiveness of their systems are all highlighted. Additionally, their claims are emphasised. Finally, a summary of the conclusions relating to the research articles that were read and examined is provided. The section's conclusion includes the cause of the detected problem.

Hritvik Gupta and Mayank Patel's experiment, published in 2021, contrasted extractive text summarization against text summarization. For example, a learning algorithm is an NLP method in which it retrieves the appropriate subject from a text material [2]. So try to complete positional encoding of topic word vectors, the proposed research summarises large amounts of textual data. In terms of efficacy, the approach described in

this study outperforms LDA topic structuring for text summarization.

Text summarization, according to Ahmed A. Mohamed and S. Rajasekaran, is a difficult task since it has so many applications [3]. This topic has been thoroughly investigated, and several solutions have been provided in the literature. The authors of this study investigate a novel meta-search-based approach. Summaries from various summarizers are specifically analysed in order to identify the best summary. The authors claim that this is the first attempt to use conceptual in a text summing situation.

As per R. S. Prasad, U. V. Kulkarni, and J. R. Prasad [4], the challenge of text summarising is mostly examined in a wide range of contexts and procedures during the last 50 years. In view of the exciting recent breakthroughs in adaptive evolving systems, this paper examines machine learning for the text summarization system. For ECTS, it employs a machine learning technique.

In 2018, V. V. Sarwadnya and S. S. Sonawane showed how labor-intensive and prone to mistake human summarization of large text volumes is. Additionally, the results of such summary might vary depending on the content [5]. The extraction concept that underlies this case study has been tested on the models in question. For both English and other foreign languages, there are several automatic text-summarizing systems available today. However, the authors note that there aren't many automatic summarizers for languages spoken in India. The formation of a computerised Marathi content summarizer seems to be the primary goal of our efforts within that domain. This article describes a multi-document marathi extractive summarizer.

Yang, D. Wen, Kinshuk, N. -S. Chen, and E. Sutinen proposed a customised text-based content summarizer in 2012 [6] to assist mobile learners in swiftly obtaining and digesting material based on their unique requirements and preferences. In this study, probabilistic language modelling methodologies are used to create a user model and an abstractive text summarising system for mobile learning that gives a personalised and automated summary. The proposed approach provides an accurate and efficient way for supporting mobile learners by summarising essential knowledge quickly and adaptively, according to trial results.

As per H. Chorfi's 2013 research, there has never been a greater need for a system that takes a text and summarises it into a brief and clear summary. Every day, there is a growing need for summarizers, particularly for those with special needs such as the elderly or the blind [7]. TS are only focusing on the features of the text rather than the author's objectives or the reader's goals. This study tackles the problem and proposes a method for acquiring implicit information. People with special needs can acquire important knowledge with the help of this technique. The writers mainly concentrate on the consequences of the implicit information transfer and the influence of the argumentative connectives.

The digital revolution has made a great quantity of data available online, but it can be difficult to discover reliable and useful data, according to C. Prakash and A. Shukla in 2014 [8]. A human can still not handle and control the amount of information that may be obtained through search engines. In order for the reader to understand the information without having to read the full page, it is vital to present it in an abstract manner. The user is then shown the produced summary, and if they approve it, it becomes the final version; otherwise, a fresh summary is created using their keywords.

Text Summarization is a kind of important advancement in natural language processing knowledge provided by S. Abujar, A. K. M. Masum, M. Mohibullah, Ohidujjaman, and S. A. Hossain [9]. The most recent recurrent neural network algorithms produce results that are noticeably better. Although significant research has already been conducted on the Bengali language, less has been done on the English language summarizer. Recognizing a text's vector representation opens the door to identifying its main ideas and evaluating how similar or different it is from other texts.

There are several approaches for blind people to interpret text, according to a 2018 argument by S. Mohan Sai, P. V. Sai Kushagra , H. S. S. S. Raviteja D, and K. Mona Teja [10]. Braille script is one example, however it is a highly inefficient method since it requires a lot of time and understanding. The writers provide a remedy for individuals who are blind since the sense of sound is unquestionably superior to the sense of touch and more accurate. In order to save time, this post outlines an effective method for distilling news items into their most important sentences. This study also covers how to convert the summarised text into voice so that blind people may also take advantage of the technology.

In the fast developing information era, A. A. P. Patil, S. Dalmia, S. Abu Ayub Ansari, T. Aul, and V. Bhatnagar remarked that there is a lot of text available right now, particularly on the web. New data are being produced consistently [11]. The ability to adequately summarise a text is vital, according to the writers, in order to maintain its clarity and content while making it simple to understand. The authors intend to develop a method that can sum up a page by altering the key text extraction using a thesaurus. Our main goal is to maintain consistency while substantially reducing a specified volume of material.

K. D. Garg, V. Khullar, and A. K. Agarwal discussed text summarization in 2021 [12]. It is a method of reducing the length of written documents without losing any of its basic context or substance. Any written piece should include a good summary that highlights the key points. This project develops a Punjabi Extractive Text Summarizer using an unsupervised machine learning approach. The usage of a method with several

modules, such as tokenizing the Punjabi text, removing stop words, generating a similarity matrix, ranking using the similarity matrix, and producing a summary, is advised.

III. Motivation

Data usage has recently become a major problem in academia, journalism, blogging, and media platforms, among other places. Because of the increase in the volume of text data, it became difficult to extract just the required information in a concise manner. To put it another way, summarising a document enables readers to locate and read the key passages. Text summarising is the process of extracting information from a source and converting it into a shorter or concise text. One of the most often used techniques is the automatic text summarizer. Large volumes of textual data are analysed using automatic text summarising software, which then distils the most crucial details into brief summaries. Software for automated text summarising is divided into two groups. Text summarization tools can be either extractive or abstract. An abstractive text summarizer approach is investigated in this study. An extractive text summarization model generates a concise summary of the material by selecting the most important phrases out from text. In order to present a final summary, this research focuses on acquiring an useful amount of data by using BERT for feature embedding, cosine similarity to compare each pair of phrases, and the Page Rank algorithm for sentence ranking.

IV. Proposed Approach

The text rank technique is one of several tactics used in NLP to achieve extractive text succinct summation. It uses sentence segmentation to identify relevant phrases in large texts, scores them using cosine similarity, and then collects the top sentences into a summary. Another method is topic modelling, which extracts important lines from large texts depending on the nature of the material. Using topic modelling and BERT large uncased to embed the phrase, we demonstrated how to extract text summaries in this work [2].

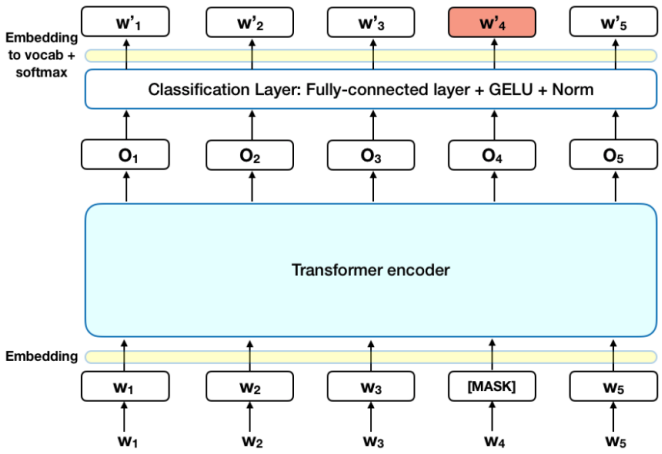

Fig. 1. Identification of BERT feature sets

To get the final score, the positional embeddings of each phrase are contrasted with those of the subjects that were extracted using LSA [2]. This is evident from Fig. 1, which shows the identification of BERT feature sets. Fig. 2 illustrates how page rank is used to condense documents.

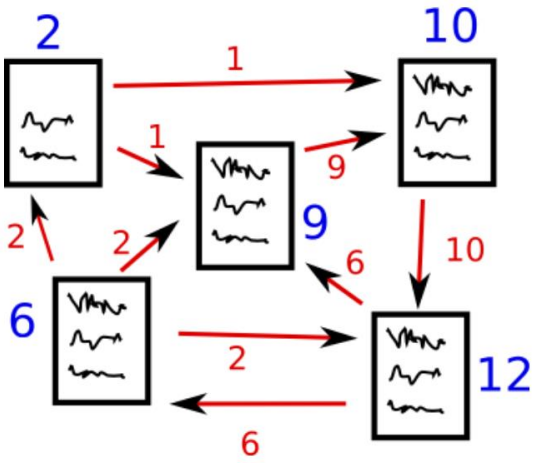

Fig. 2. Use of Page Rank to summarize documents

The block design for the Page Rank flow for the summarization process is shown in Fig. 3. Here, the process starts with the articles that need to be distilled. The text from these articles is concatenated, then it is extracted. The sentences are then formed by dividing the texts. The vectors of these phrases are then discovered. Then, these vectors are created and evaluated against a similarity matrix. After that, a graph is generated using the similarity matrix. Calculated and organised are the sentence rankings. The articles' summaries are then acquired.

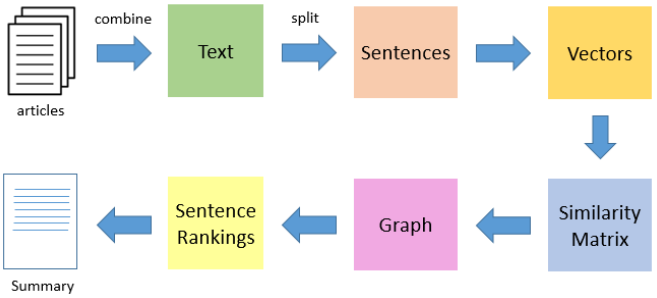

Fig. 3. Flow of the Page Rank Model for summarization process

Now, the complete working of our project can be viewed with the help of flowchart shown in Fig. 4.

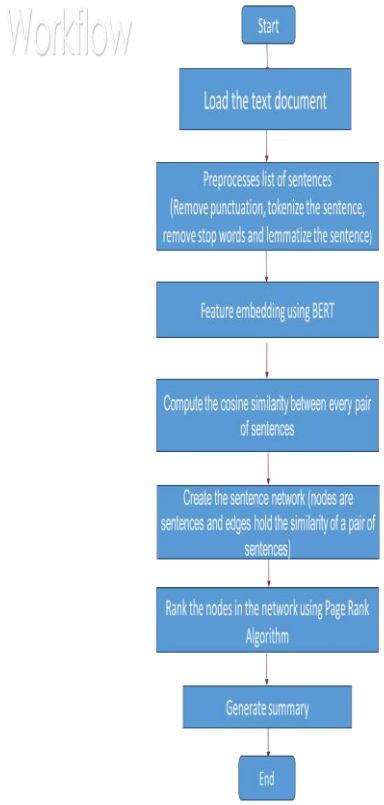

Fig. 4. Workflow of proposed model

The loading of the text file that has to be summarised starts the procedure. The sentences in this text document are fixed during preparation. Lemmatize the sentences, remove all punctuation, rank and tokenize the sentences, remove all stop words, and more. Now, we're utilising BERT to incorporate the functionalities. The cosine similarity between each pair of phrases are now calculated. Now, graph theory is applied to a network of texts. We are aware that a graph has nodes and edges. The sentences are represented here as nodes on a graph, with edges indicating how similar two phrases are to one another. The Page Rank Algorithm is now used to rank the pages. The summary is then produced.

Additionally, Fig. 5 helps to illustrate the BERT and Page Rank model-based summary creation process. The text passes through preprocessing and correction, as can be seen. A sentence similarity matrix is created once the vectors have been extracted. Using the graph theory, the network of texts and similarity index are created. Using the Page Rank Algorithm, the sentences are ranked, and then a summary is produced.

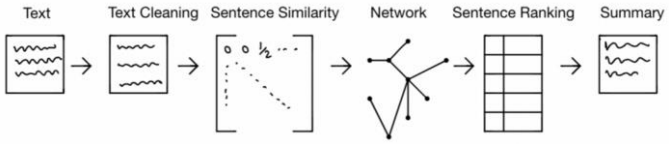

Fig. 5. Process of summary generation using BERT and Page Rank model

V. RESULTS

This platform was built using the Spyder console's anaconda operating mode and the Python language. NLTK for text processing, Panda for dataset acquisition, Numpy, and Sklearn's cosine similarity are also included. The BERT trained network is accumulated using Tensor-flow Hub. Once all of the libraries have been included, the data set must be imported; for summarising, the news summary input data from Kaggle datasets was utilised. Each story in the collection has a headline, complete text, summary text, and links to more news pieces.

In this experiment, the first 100 articles from the news summary collection were used. Use the Rouge scorer optimization method to study its reactions with both the new and old summary. Table 1 shows the estimated amount Rouge-1 and Rouge-L scores for the first 100 documents, 50 documents, and 30 documents for integrative text summarization utilising BERT Integration..

Table-1: Average Rouge score of first 100, 50 and 30 documents

No. of documents	Rouge-1			Rouge-L		
	F	P	R	F	P	R
100	0.35	0.82	0.23	0.34	0.79	0.22
50	0.32	0.74	0.21	0.3	0.7	0.19
30	0.31	0.75	0.2	0.29	0.7	0.19

Rouge-1: The overlap of unigrams between the system summary and reference summary is indicated by the notation Rouge-1.
Rouge-L: It leverages the LCS to find which word has the largest matching sequence. Adopting LCS gives the benefit of just required in-sequence matches that accurately reflect sentence-level word order rather than consecutive matches. Because the threshold in-sequence typical n-grams are auto involved, no precise n-gram length is required..

Precision: Accuracy is a big assistance in determining how relevant or valuable the system-generated summary is. The fraction of words that overlap is calculated as a percentage of all words. In the automatically created summary.

The proportion of words which divide across in relation to the total number of words in the network.

The percentage of the reference summary that the system summary is able to retrieve or capture in the context of ROUGE is referred to as recall.

Recall is calculated as the proportion of terms that overlap in the reference summary.

The **F-measure** connects recall and precision and represents the system's accuracy.
F-Measure = 2 x Recall x Precision / Recall + Precision

Figures 6, 7, and 8 show the Rogue 1 metrics for 30 documents, 50 documents, and 100 documents, respectively.

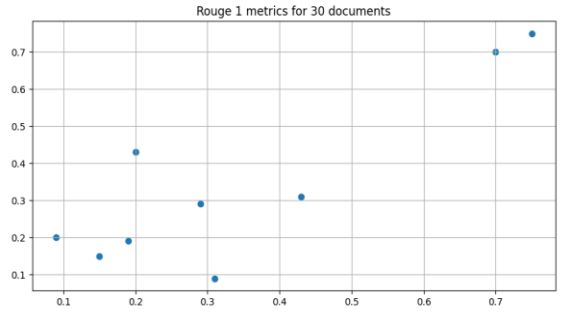

Fig. 6: Rogue 1 metrics for 30 documents

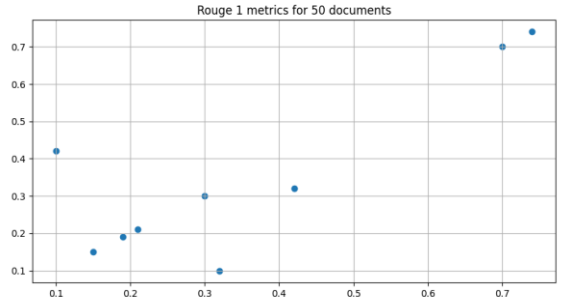

Fig. 7: Rogue 1 metrics for 50 documents

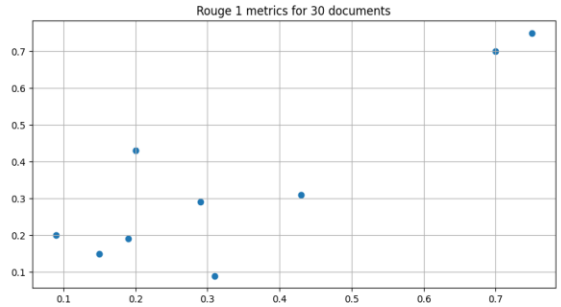

Fig. 8: Rogue 1 metrics for 100 documents

V. Conclusion

In this paper, we have analysed and contrasted the literature that has already been written about several text summarization tools. The present techniques are systematically grouped based on the fundamental ideas employed in their procedures. Following a review of the literature, we have suggested our approach, which addresses the problems. This work is crucial for the comparative analysis of different text summarizer methodologies, which is a requirement for resolving the problems with document abstraction. We combined NLTK with BERT. We have put into practise the concept of easily and quickly summarising a text. On both the produced summary and the actual summary, we used the Rouge scorer approach to evaluate the outcomes. the 50, 30, and 100 papers with the highest average Rouge scores. The outcomes were encouraging.

References

[1] H. Gupta and M. Patel, "Study of Extractive Text Summarizer Using The Elmo Embedding," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 829-834, doi: 10.1109/I-SMAC49090.2020.9243610.

[2] H. Gupta and M. Patel, "Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 511-517, doi: 10.1109/ICAIS50930.2021.9395976.

[3] A. Mohamed and S. Rajasekaran, "A text summarizer based on meta-search," Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005., 2005, pp. 670-674, doi: 10.1109/ISSPIT.2005.1577177.

[4] R. S. Prasad, U. V. Kulkarni and J. R. Prasad, "Machine learning in Evolving Connectionist Text Summarizer," 2009 3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication, 2009, pp. 539-543, doi: 10.1109/ICASID.2009.5277001.

[5] V. V. Sarwadnya and S. S. Sonawane, "Marathi Extractive Text Summarizer Using Graph Based Model," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697741.

[6] G. Yang, D. Wen, Kinshuk, N. -S. Chen and E. Sutinen, "Personalized Text Content Summarizer for Mobile Learning: An Automatic Text Summarization System with Relevance Based Language Model," 2012 IEEE Fourth International Conference on Technology for Education, 2012, pp. 90-97, doi: 10.1109/T4E.2012.23.

[7] H. Chorfi, "Get only the essential information: Text summarizer based on implicit data," Fourth International Conference on Information and Communication Technology and Accessibility (ICTA), 2013, pp. 1-4, doi: 10.1109/ICTA.2013.6815299.

[8] C. Prakash and A. Shukla, "Human Aided Text Summarizer "SAAR" Using Reinforcement Learning," 2014 International Conference on Soft Computing and Machine Intelligence, 2014, pp. 83-87, doi: 10.1109/ISCMI.2014.22.

[9] S. Abujar, A. K. M. Masum, M. Mohibullah, Ohidujjaman and S. A. Hossain, "An Approach for Bengali Text Summarization using Word2Vector," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944536.

[10] K. Mona Teja, S. Mohan Sai, H. S. S. S. Raviteja D and P. V. Sai Kushagra, "Smart Summarizer for Blind People," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), 2018, pp. 15-18, doi: 10.1109/ICICT43934.2018.9034277.

[11] A. P. Patil, S. Dalmia, S. Abu Ayub Ansari, T. Aul and V. Bhatnagar, "Automatic text summarizer," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2014, pp. 1530-1534, doi: 10.1109/ICACCI.2014.6968629.

[12] K. D. Garg, V. Khullar and A. K. Agarwal, "Unsupervised Machine Learning Approach for Extractive Punjabi Text Summarization," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), 2021, pp. 750-754, doi: 10.1109/SPIN52536.2021.9566038.