IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Review of Text Summarization Techniques of Documents

[1] SHUBHAM U. PAWAR, [2] OM S. BEHARE, [3] SUMIT D. UMAP, [4] AKSHAY K. ADHAV, [5] BHUSHAN B. PAWAR, [6] RAM S. THAKARE, [7] PROF. HARSHADA M. RAGHUWANSHI

[1,2,3,4,5,6] U.G. Students, Department of Computer Science & Engineering, DRGITR, Amravati,
[7] Assistant Professor, Department of Computer Science & Engineering, DRGITR, Amravati,

Abstract: This paper presents the analysis and comparison of existing literature relevant to various text summarization of document and the mechanisms associated with it. Though, the literature consists of a lot many research contributions, but, here, we have critically and exhaustively analyzed recent research and review papers that are pertinent to text summarization of document systems. Based on the basic concepts used in their mechanisms, the existing approaches are categorized. The emphasis is on the concept used by the concerned authors, the methodology used for experimentations and the performance evaluation parameters. The claims of the researchers are also highlighted. Our findings from the exhaustive literature review are mentioned along with the identified problems. This paper is very important for the comparative study of various text summarizers approaches which is prerequisite for solving related issues.

Index Terms – Artificial Intelligence, Chatbots, Health Care Systems, Machine Learning, Natural Language Programming

Introduction

With the expansion of social media, online education services, and all professional fieldwork during the past several years, knowledge on the internet has grown at an incredible rate. Textual data is the primary format in which this information is expanding. Therefore, processing and comprehending one such vast amount of data has emerged as the main challenge. The only way to solve this issue is to condense the textual material, also known as summarising [2], which is the process of reducing the amount of the textual data. It is difficult to condense material into concise summaries, though. It demands having a thorough knowledge of the text that will be summarised. One NLP strategy, text summarization, which is covered later in this section, can be used to solve this issue [1].

NLP has become more crucial as a result of the daily growth in the amount of text data created. Every time a person uses the internet, data is created and saved, making the capacity to make data-driven decisions essential for every organisation. Automatic text summarizer and Topic Modelling are two tools in the field of natural language processing that are frequently used to reduce the one such vast quantity of data over the data centres. It aids in data structure by reducing the amount of the data while keeping the pertinent information.

The use of text summarization on the internet has grown recently. Numerous people use text summarization. They might be regular internet users looking for news, e-learners seeking specialised educational resources, scientists researching certain works, or anyone with special needs like the blind or senior citizens seeking concise and clear information. By condensing a significant quantity of information into a summary, text summarization can speed up consumers' access to the information they need. There is general agreement, however, that summarization—the process of condensing a huge amount of information into a summary while retaining just the most important details—is an extremely challenging task. Similarly, the method of human summarising is trying to comprehend, analyse, and synthesize a given content before creating a new document that serves as its summary.

I. LITERATURE REVIEW

This section presents the critical analysis of existing literature which is relevant to text summarizer systems and the mechanisms associated with it. Though, the literature consists of a lot many research contributions, but, here, we have analysed most recent, relevant and pertinent research and review papers. The existing approaches are categorized based on the basic concepts involved in the mechanisms. The emphasis is on the concepts used by the concerned authors, the platform used for experimentations and the performance of systems. Their claims are also highlighted. Finally, the findings are summarized related to the studied and analysed research papers. Section concludes with the motivation behind identified problem.

In 2021, Hritvik Gupta and Mayank Patel presented an experiment in 2021 that contrasted with extractive text summarization for text summarization. Subject modelling, on the other hand, is an NLP activity that pulls the pertinent topic from the textual source [2]. One such technique selects all the pertinent themes from the text using Latent Semantic Analysis (LSA) with reduced SVD. The proposed research summarises lengthy textual documents using LSA topic modelling, TFIDF keyword extractors for each sentence, and BERT encoder models for encoding the sentences from textual documents in order to retrieve the positional embedding of topics word vectors. The method suggested in this research is capable of outperforming text summarization using Latent Dirichlet Allocation (LDA) topic modelling in terms of performance.

Since text summarization has so many uses, Ahmed A. Mohamed and S. Rajasekaran argued in 2005 that it is a significant challenge [3]. Numerous ways have been suggested in the literature to address this issue, which has been the subject of in-depth study. In this study, the authors examine a novel meta-search-based strategy. To get the best summary, in particular, summaries from different summarizers are compared. This is the first effort that, as far as the authors are aware, uses meta-search in a text summarization context. In these tests, the authors used 5 different summarizers and data from the DUC-2002.

R. S. Prasad, U. V. Kulkarni, and J. R. Prasad talked about how throughout the past 50 years, the issue of text summarization has been approached from a variety of angles, in a variety of domains, and using a variety of paradigms [4]. With exciting new advancements in adaptive evolving systems in mind, this research aims to examine machine learning for the text summarization system. This study focuses on the machine learning method for an Evolving Connectionist Text Summarizer ECTS, which is a Computational Intelligence (CI) system that operates in real time and adapts its structure and functionality via constant interaction with the environment and other systems.

In 2018, V. V. Sarwadnya and S. S. Sonawane demonstrated how human summarising of huge text volumes is time-consuming and error-prone. Furthermore, the outcomes of such summarization may yield diverse results for a given text [5]. This case study is based on an extraction idea that has been executed on the models under consideration. Today, there are several automated text summarising systems available for English and other foreign languages. However, when it comes to Indian languages, the authors see an insufficient number of automated summarizers. Our efforts in this area are primarily focused on building an automated text summarizer for the Marathi language. The authors are excited to use the ROUGE measure to analyse the obtained summary. A multi-document marathi extractive summarizer is described in this work.

In order to help mobile learners quickly retrieve and process information based on their interests and preferences, Yang, D. Wen, Kinshuk, N. -S. Chen, and E. Sutinen introduced a personalised text-based content summarizer in 2012 [6]. A user model and an extractive text summarising system are built using probabilistic language modelling approaches in this study to produce a tailored and automated summary for mobile learning. According to experimental findings, the suggested solution offers a correct and effective method for assisting mobile learners by summarising crucial knowledge swiftly and adaptively.

The demand for a tool that takes a text and condenses it into a quick and succinct summary has never been higher than it is now, according to H. Chorfi's 2013 article. The need for summarizers is becoming more and more urgent every day, especially for persons with special requirements like the blind or the elderly [7]. The majority of Text Summarizers (TS) handle the original material by compressing it, which inevitably results in information loss. TS are simply concentrating on the text's characteristics, not on the author's intentions or the purpose of the reader. This research addresses this issue and introduces a system that focuses on gaining implicit knowledge. Such a method aids in the acquisition of crucial information by persons with special needs. The authors focus primarily on the implications of the argumentative connectives' implicit information transmission and their impact on.

Due to the digital revolution, a vast amount of data is available online, but finding accurate and pertinent data is not a simple process, according to C. Prakash and A. Shukla in 2014 [8]. The amount of information that can be found through search engines is still much more than a person can handle and manage. Therefore, it is necessary to convey the information in an abstract manner so that the reader may quickly get the meaning without having to read the entire page. In this study, the human-aided text summarizer "SAAR" for a single document is proposed. As a result, the generated summary is displayed to the user, and if they accept it, it becomes the final version; otherwise, a new summary is designed based on their keywords. The performance of the suggested strategy compares quite strongly with other techniques in terms of precision, recall, and F-score, according to the results of testing using DUC2006 documents.

One of the notable research fields in Natural language processing in 2019 is Text Summarization, which was developed by S. Abujar, A. K. M. Masum, M. Mohibullah, Ohidujjaman, and S. A. Hossain [9]. The most recent recurrent neural network techniques deliver significantly improved outcomes. While some notable study has previously been done for the Bengali language, not as much has been done for the English language summarizer. Word2vector is just one of several requirements for data analysis purposes. Being able to recognise a text's vector representation paves the way for determining its core themes and for assessing how similar or distinct it is to other texts. The top-ranked sentences and words may be quickly determined using the word2vector-generated matrix, whether they are in generic or domain-specific form. In the context of Bengali text summarization, a word2vector technique has been discussed in this study.

There are several ways for blind persons to read text, according to a 2018 discussion by K. Mona Teja, S. Mohan Sai, H. S. S. S. Raviteja D, and P. V. Sai Kushagra [10]. One of the examples is Braille script, however it is a very ineffective technique since it takes a lot of time and skill. In light of the fact that the sense of sound is definitely superior to the sense of touch and more precise, the authors provide a solution for those who are visually impaired. This essay discusses a useful technique for condensing news articles into essential phrases in order to avoid having to read the entire content each time. In this study, many methods, including Luhn's Algorithm, Latent Semantic Analysis Algorithm, and Text Ranking Algorithm, have been described in depth and implemented. Additionally, this study discusses how to turn the summarised text into speech so that blind persons may also benefit from the technology.

A. A. P. Patil, S. Dalmia, S. Abu Ayub Ansari, T. Aul, and V. Bhatnagar noted how there is a lot of text available now, particularly on the web, in the rapidly expanding information era. There is a steady production of new data [11]. The authors find that it is crucial to be able to effectively summarise a document in order to retain its meaning and clarity while making it easy to grasp. Using a thesaurus to try to alter the key text extraction, the authors want to create an approach that can summarize up a page. Our fundamental objective is to preserve consistency while drastically shrinking a given volume of material.

Text summarising was described by K. D. Garg, V. Khullar, and A. K. Agarwal in 2021 [12]. It is the technique of condensing text documents while maintaining their general context and substance. Any written work should include a decent summary that illustrates the main ideas. A crucial use of Natural Language Processing (NLP), text summarising employs several NLP technologies to extract useful information from provided text. In this study, an unsupervised machine learning technique is used to create a Punjabi Extractive Text Summarizer. It is suggested to use a process that consists of many modules, including tokenizing the Punjabi text, eliminating stop words, creating a similarity matrix, ranking using the similarity matrix, and creating a summary.

II. Motivation

Data overuse has recently emerged as a serious issue in the fields of education, journalism, blogging, social media, etc. It became difficult for a human to extract only the important information in a compact form due to the growth in the volume of text data. In other words, summarizing a document makes it possible for others to find and read the important and valuable texts. Text summarization is the process of taking information from a document and turning it into a brief or succinct text. Automatic Text summarizer is one of the main strategies that is commonly utilized. Automatic text summarization software analyses enormous amounts of textual data and condenses it into concise summaries that include the material's most important information. Separating automatic text summarization software into two categories There are two types of text summarizers: (1) extractive (2) abstract. The extractive text summarizer technique is sought for in this paper. By selecting the most pertinent phrases from the text document, an extractive text summarization model creates a brief summary of the material. In order to construct the final summary, this study focuses on obtaining the valuable quantity of data utilizing BERT for feature embedding, cosine similarity to compare each pair of sentences, and Page Rank algorithm for sentence ranking.

III. Proposed Methodology

There are numerous methods for performing extractive text succinct summation in NLP, including the text rank method, which involves extracting pertinent sentences from lengthy texts using sentence embedding, scoring them using cosine similarity, and combining the top sentences into a summary. Another method is topic modelling, which extracts pertinent sentences based on the subject matter of lengthy texts. In this study, we showed how to extract text summaries utilising topic modelling and BERT big uncased to embed the sentence [2].

BERT (Bidirectional Encoder Representations from Transformers). It is the deep learning model that has been pre-trained using unlabeled text. created and released by Google AI language researchers. BERT is a model that combines the attention model with a bidirectional layer for language modelling. BERT employs an attention mechanism that recognises the contextual relationships among the text's words. Contextual encoding of phrases in lengthy textual documents is typically included in BERT models. The long textual material in this research project was encoded using the same contextual encoding method as BERT. The experiment that will be used to create text summaries utilizing an extractive text summarizer technique has been shown in this study. We're going to use BERT to create position word embeddings for all of the topic word vectors from the document's LSA topic modelling and the extracted keywords from the TFIDF vectorizer. The positional embeddings of each phrase are compared to those of the topics retrieved using LSA to get the final score [2]. This can be seen from Fig. 1 wherein the Identification of BERT feature sets can be viewed. The usage of page rank to summarize documents can be seen in Fig. 2.

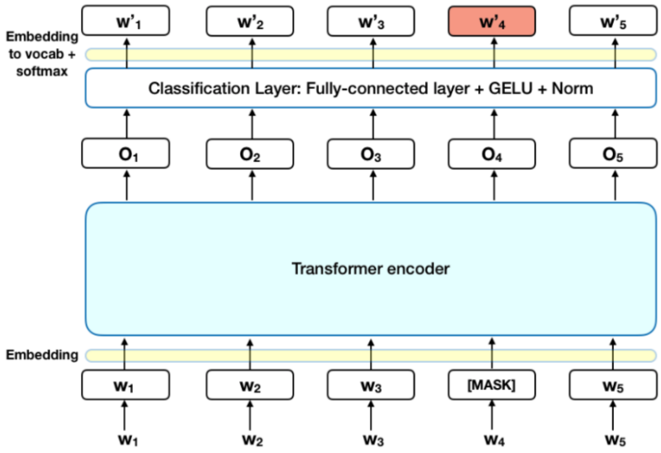

Fig. 1. Identification of BERT feature sets

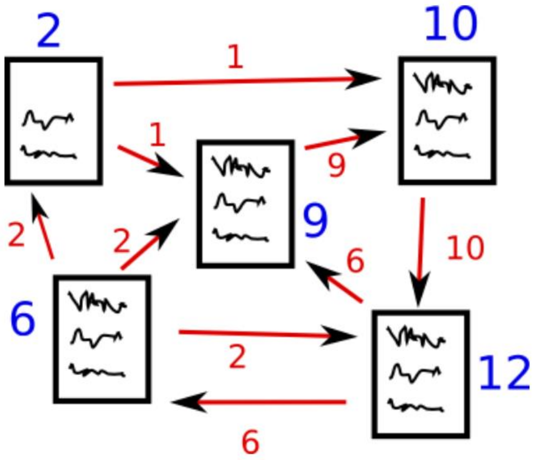

Fig. 2. Use of Page Rank to summarize documents

Fig. 3 shows the block diagram representing the flow of Page Rank for summarization process. Here the process begins with articles which are to be summarized. These articles are combined and the text from the articles are extracted. The texts are then splits to form the sentences. We then find the vectors of these sentences. These vectors are then formed and compared to similarity matrix. A graph is then plotted from the similarity matrix. The sentence rankings are calculated and arranged. Finally, the summary of the articles is obtained.

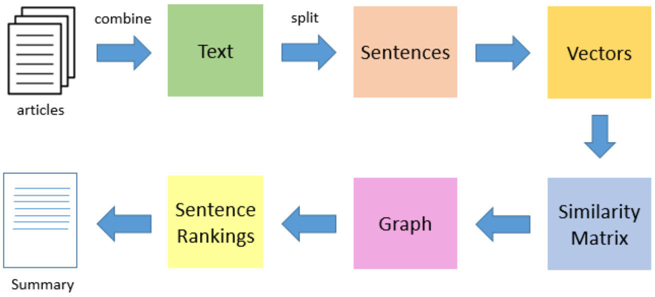

Fig. 3. Flow of the Page Rank Model for summarization process

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

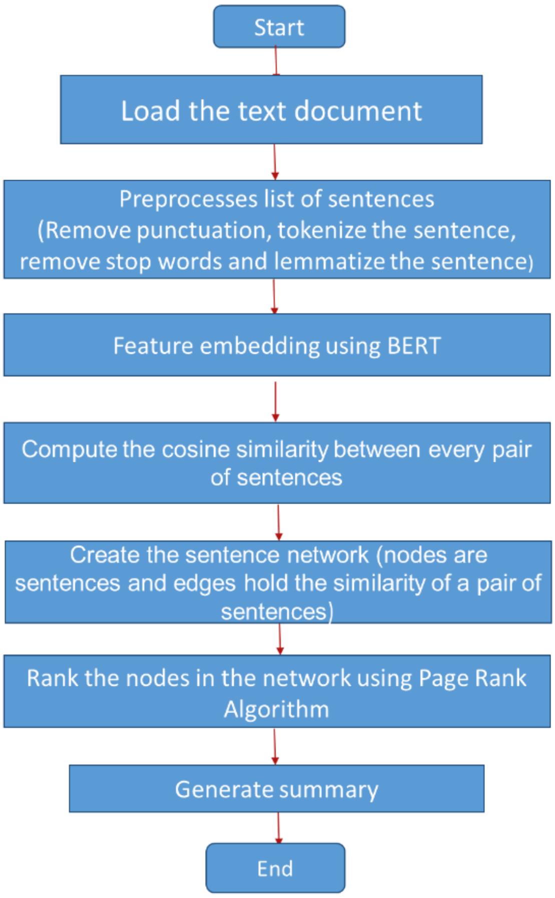
Start
Load the text document
Preprocesses list of sentences
(Remove punctuation, tokenize the sentence, remove stop words and lemmatize the sentence)
Feature embedding using BERT
Compute the cosine similarity between every pair of sentences
Create the sentence network (nodes are sentences and edges hold the similarity of a pair of sentences)
Rank the nodes in the network using Page Rank Algorithm
Generate summary
End

Fig. 4. Workflow of proposed model

The process begins with loading of the text document which is to be summarized. This text document undergoes the preprocessing wherein the sentences are rectified. This includes removal of punctuation, ranking and tokenization of sentences, removal of stop words and lemmatize the sentences. Now we embed the features using BERT. We now compute the cosine similarities between every pair of sentences. Now, a network of sentences is created and graph theory is utilized. We know that a graph consists of nodes and edges. Here, the sentences are mapped onto a graph, where nodes are the sentences and the edges hold the similarity of a pair of sentences. Now the pages are ranked using the Page Rank Algorithm. Finally, the summary is generated.

Fig. 5 also assists to visualize the process of summary generation using BERT and Page Rank model. As it can be seen, the text undergoes the preprocessing and rectification. The vectors are extracted and a sentence similarity matrix is formed. The network of sentences and similarity index is formed using the graph theory. The sentences are ranked using Page Rank Algorithm and finally the summary is obtained.

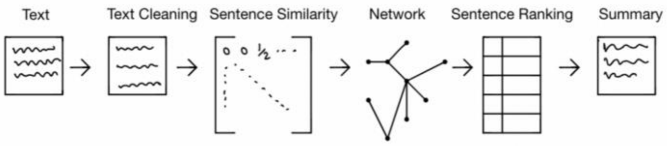

Fig. 5. Process of summary generation using BERT and Page Rank model

IV. Results

This program is created using the python programming language and written in the Spyder console using the anaconda environment. The packages that are used is NLTK for text processing, Tensor-flow hub for downloading the BERT pretrained model, cosine similarity from the Sklearn library and other common libraries like a panda for importing the dataset, Numpy etc. Once all the libraries are imported the next step is to import the data set, for the purpose of summarization the news summary data set have been used from the Kaggle datasets . The dataset contain headlines, complete text, summarized text , and news articles link for each article in the dataset. In this experiment, the first 100 documents have been executed from the news summary dataset. And evaluate the result by using the Rouge scorer method on the generated summary and the original summary.

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.