JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text Summarization: Providing a Summary of any given input in a different language

Mrs. Vishakha Shelke
Assistant Professor
Department of Computer Engineering
Universal College of Engineering
Vasai, India
vishakha.shelke@universal.edu.in

Mr. Ritik Pandey
Department of Computer Engineering
Universal College of Engineering
Vasai, India
pandeyritik010@gmail.com

Mr. Dhiren Parekh
Department of Computer Engineering
Universal College of Engineering
Vasai, India
dhirenparekh7@gmail.com

Ms. Nutan Thakare
Department of Computer Engineering
Universal College of Engineering Vasai, India
nutanthakare13@gmail.com

***Abstract*—At this present time where daily information is produced on the internet. With this, the quantity of information in the form of documents, articles, blogs, etc is increasing daily. The international data corporation analyzed that the total amount of data moving across the globe would increase from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025. An average person spends 30 to 40 minutes reading news or getting information from the internet. So it is necessary to get summarized information from the original data while containing the most important sentences. By this, a lot of reading time can be saved. Automatic text summarization is the process of creating a shorter form of original data using a machine. We are taking the input in text and pre-processing it by removing blank space etc. After that different token is allocated to different sentences. These tokenized sentences are further given the sentences vector to generate weights of words using the TF - IDF algorithm. Lastly, the sentences with the higher score will be taken out as a summary.**

Keywords— Text summarization, Natural Language Processing, TF-IDF.

I. INTRODUCTION

In this digital era, daily data produced on the internet is huge in the amount which will increase to 180 zettabytes in 2025 and nowadays most people get information from the internet source having similar sentences and in a bigger size which results in a lot of wastage of prestigious time.

An automatic text summary generator is the tool that helps to produce a summary. In this informative era, it becomes essential to use modern technology to solve our problems. This proposed system is going to use natural language processing (NLP) and rank the sentence according to the TF-IDF value for generating a content summary. First, the input will be given in form of text to the system and passed to the pre-processing phase where cleaning is done like removing capital, punctuation mark, and numbers using. This is done to make the text in only word format. Next, this cleaned text is used to make sentences that will be tokenized by giving tokens to each sentence which will break up each sentence into words. This all process will be done by using the TF-IDF algorithm and according to the score of the sentences and by this score the important sentences can be taken out as summary output.

The rest of the paper is structured as follows: Section II reports on some of the recent works related to the topic. Section III describes the proposed system. Section IV shows the results of the experiment and Section V contains the conclusion

II. LITERATURE SURVEY

In the paper presented by Aditya Jain et al have worked on extractive text summarization using word vector embedding. In this, the input document is broken down into sentences and the feature extraction is performed over these sentences and fed into the neural network for training and prediction. The neural network is used to decide the inclusion of the sentence in the summary. And the sentence score of the feature is calculated by taking the mean of each dimension of all the word vectors, forming a vector. Testing of this project has been performed on DUC 2002 dataset, where up to 284 documents were used in various test experiments. ROUGE scores (1, 2, and L) and the results are effective. [1]

In the paper presented by Li-Jyun Chen, Chih-Ying Chen, Hung-Yu Chen, Guan-Yu Chen & Yen-Wen Chen have proposed the system first applies the text frame detection scheme to identify the text parts of the document and utilizes the Optical Character Recognition (OCR) to convert the text image into a pure text file. There project aimed to implement the automatic text summarization algorithm to get the brief meaning of the whole text. And with the help of their project visually-impaired people can listen to the summary of the document through the existing text to voice tool. They have also used the system of neural network model which is adapted to find out text blocks before recognizing the text and to optimize the Chinese automatic text summarization system. [2]

The paper presented by J.N. Madhuri et al has done a demonstration on a single document to perform a statistical method on extractive text summarization. This method of extraction of sentences, which gives the idea of the input text in a short form, is presented. They have implied this method in which the sentences are ranked by assigning weights and they are ranked based on their weights. Sentences that are highly ranked are extracted from the input document as it extracts important sentences which direct to a high-quality summary of the input document These documents are stored as a summary and later it converts the same summary into an audio file i.e mp3 format. [3]

The paper presented by Sagarika Pattnaik and Ajit Kumar Nayak et al has done text summarization using clustering and cosine similarity which segregates sentences. The model adopts an extractive method for summarizing the Odia text document. Odia is a morphologically complex language still the proposed system can make a summary. The objective of the model is to produce a precise and informative summary with minimum redundancy. As there are no NLP tools developed for the Odia language, the suggested technique has proven to be robust and effective. A cosine matrix table is used to sort similar sentences in one cluster. For evaluation, they have taken into consideration 20 humans. After evaluation, the proposed model has got a satisfactory result by scoring an average score value of 60.272%. [4]

In the paper presented by Kasimahanthi Divya, Kambala Sneha, Bassetti Sowmya, and G Sankara Rao have proposed Text summarization using natural language processing techniques and using algorithms like page rank algorithms, etc. While those algorithms satisfy the goal of textual content summarization, they can't generate new sentences which aren't withinside the record like humans. While using this project they faced some grammatical errors, which were solved using Deep Learning. The use of deep getting to know builds green and speedy version for textual content summarization. The use of deep learning methods also helps them to generate summaries that can be formed with new phrases and sentences and also which are grammatically correct. [5]

In the paper presented by Anish Jadhav, Rajat Jain & Steve Fernandes have described an algorithm based on a combination of both Extractive Summarization and Abstractive Summarization approaches. Initially, crucial sentences are diagnosed and stitched collectively to shape a consolidated report. The importance of a sentence is selected in mild measurable and semantic highlights of sentences. And this shorter representation is passed through an Encoder-Decoder model to generate a concise summary representing the whole article, which is semantically and linguistically correct, by understanding the whole content and reletting it in its own words. This proposed methodology focuses only on the relevant sentences and passes them to the Bi-Directional RNN for identifying and representing the core idea of the article. [6]

In the paper presented by Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh and Ayoub Bagheri have proposed a query-oriented text summarization technique by extracting the most informative sentences. In this, some features are extracted from the sentences, each of which evaluates the importance of the sentences from an aspect. In this paper, eleven first-rate capabilities are extracted from each of the sentences. This paper has shown that the use of more suitable features leads to improved summaries generated. To evaluate the automatically generated summaries, the Recall Oriented Understudy for Gisting Evaluation (ROUGE) criterion has been used. They also used the features for summarizing DUC 2007 corpus. To enhance the performance of the summarization, they have used some query-dependent features that lead to higher ROUGE values. [7]

In the paper presented by Meena S M, Ramkumar M P, Asmitha R E and Emil Selvan G SR have used the TFRSP (Text Frequency Ranking Sentence Prediction) algorithms to generate a particular precis that makes use of supervised and unsupervised mastering algorithms. In this project, they have used the combination of TF-IDF-TR (Term Frequency – Inverse Document Frequency – Text Rank) as an unsupervised learning algorithm and the Seq2Seq (Sequence to Sequence) model as a supervised learning algorithm to obtain the benefits of both extractive and abstractive summarization. And the results of the proposed TFRSP approach are compared with the existing methods of text summarization using the Recall Oriented Understudy for Gisting Evaluation (ROUGE) and attains a high ROUGE score, and according to the scores, the summary can be achieved more accurately. [8]

The paper presented by Narendra Andhale and L.A. Bewoor states that a comprehensive survey on both extractive and abstractive approaches in text summarization. Both extraction and abstraction yield good results depending on the context. But in Extraction based summaries they found that sentences are naturally longer than average. The summary may sometimes include unimportant information. In this, the Summary texts suffer from a lack of flow, as extracted contents are taken from different parts of the document leading to sudden shifts in topic, and in abstractive based summary they found that the quality of abstractive summary depends on the deep linguistic skill. Abstractive summaries often fail to capture the semantic relationship between important terms in the document. The quality of abstractive summaries is dependent on deep linguistic skills. [9]

This paper presented by Neelima Bhatia and Arunima Jaiswal has investigated the popular and important work done in the field of single and multiple document summarizations, generous distinctive prominence towards pragmatic approaches, and extractive techniques. They saw that this subject requires a huge amount of optimal work, and earlier attempts have ranged from summarizing scientific words to newscast stories, automated mail letters, announcements, and blogs. They noticed that self-possessed extractive and abstractive techniques were still being used, even though the request on arrow had been made. They discovered a difference between single document and multi-document summarization. Following that, some motivating efforts have been made up to now within earlier research, which they found to be of good scope for future research; additionally, these works focus only on trivial minutiae associated with a broad-ranging summarization progression, rather than proposing an all-encompassing summarization scheme. [10]

This paper presented by Prabhudas Janjanam and CH Pradeep Reddy surveyed different methods and algorithms associated with automatic summarization. In their study, they focused on the representation of features, sentence selection, and summary generation using machine learning and recent graph and evolutionary methods. They studied previous and recent literature on models and algorithms relating to automatic summarization. As part of this paper, they discuss ML techniques and methods to perform extractive and abstractive summarization using a graph, semantic-based, and optimization. In their study, they observed that most of the summarization tasks addressed by researchers are domain-independent. In their study, they used mostly news-specific datasets. There is a possibility that low-significant sentences can be included in the summary, thus wasting space. Also, they found that sentences that are identified as salience may not be extracted for a summary generation. As they noted, the semantic approach needs to be improved for depicting the sentiment of text since a syntactic approach proved to be effective for summarizing text. [11]

In this paper presented by Adhika pramita widyassri, Ahmad zainul fanani, Abdul syukur,ruri suko basukiEdy, and Noersasongko they have identified and analyzed methods, datasets, and trends in automatic text summarization research from 2015 to 2019. They have used a systematic literature review (SLR) method for automatic text summarization. They found that the extractive approach is still quite popular in the past three years. Because the extraction approach is easier than the abstractive and the opportunity to combine methods still exists, they conclude that the text summarization research trend has undergone a slight shift in the last three years, where new things have emerged that are trends that are leading to optimization. In their analysis, they found that the technique used is an extractive technique for improving accuracy by optimizing an existing method. [12]

III. PROPOSED SYSTEM

In Fig 1 We show the proposed system and explore the different modules involved along with the various models through which this system is understood and represented. There is an incredible want to lessen an awful lot of this newsletter's facts to shorter, targeted summaries that seize the salient details, each so we can navigate it greater effectiveness as well as check whether the larger documents contain the information that we are looking for. The main idea behind automatic text summarization is to find a short subset of the most essential information from the entire set and present it as a result Automatic textual content summarization techniques are significantly hard to cope with the ever-developing quantity of textual content statistics to be had online and offline. For text summarization, the task system required high accuracy to make the model effective and accurate since TF-IDF has higher accuracy than other algorithms and it also handles large data sets which are required for summarizing the text.

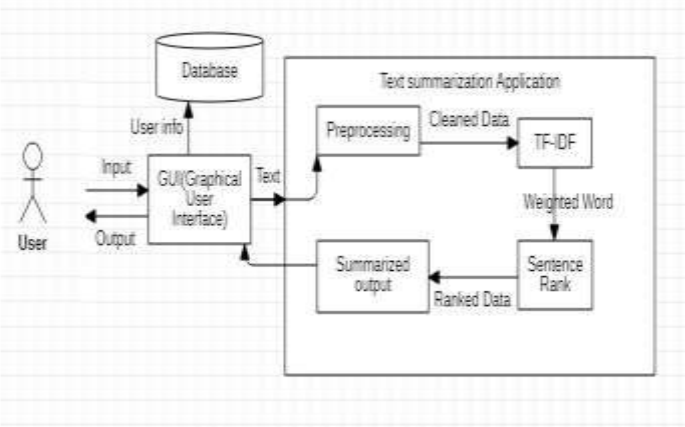

Fig 1. System Architecture

The performance parameters such as time for execution, accuracy, agent, reliability, performance, and specificity of the TF-IDF are high thus making it a good option for the text summarization process. The modules of our project are as follows:

A. *Multi-Lingual*

A person who can speak and understand more than four languages is called a Multilingual. In our project, we have used Google Translate to generate a summary of different languages. We can generate the output in 4 different languages which are English, Hindi, Marathi, and Gujarati. So if the input is in English then it will go directly to module B. But if the input is in Hindi, Marathi, Gujarati then it will translate to English using the google trans python library then it will go to module B.

B. *Pre-processing*

Text pre-processing is an important step to generating accurate output of the given. It is a process in which cleaning of the raw text is done which contains some stop words removal, space removal, and stemming.

C. *Tokenization*

There are a lot of tokens that are not worthy of content such as question marks, surprise marks, commas, and so on. For this purpose, in the tokenization process, these meaningless tokens are deleted from the text and the sentences are broken down into meaningful tokens. These significant tokens are separated from every different way of means of the white area or punctuations.

D. *TF-IDF*

It is an algorithm that uses a numerical statistic that is intended to highlight how important a word is to a document. How many times a word appears in a document and the Inverse document frequency of the word across a set of documents.

E. *Scoring the sentence*

Using the TF-IDF values of each sentence we can find the topmost sentence. Therefore, we first sort the sentence according to its value in descending value, and then we fetch the topmost sentence to produce a summary.

F. *Summary*

Here if the input was given in English then the output will be shown as it is. But if the input was given in Hindi, Marathi, or Gujarati then the produced summary is translated back to its input language.

IV. RESULT AND DISCUSSION

In table 1. Shows the comparison between the Rouge values of the Existing algorithm for TF IDF. F-measure, Precision, and Recall are the parameters that serve for comparing the actual purposed algorithm.

TABLE 1. Performance Comparison of existing algorithms with proposed TF IFD algorithm using ROUGE score

Algorithms	ROUGE VALUE		
	F-measure	**Precision**	**Recall**
Bert	0.052567	0.028564	0.076714
Word Vector	0.126349	0.132689	0.109128
TF IDF Method	**0.248723**	**0.289640**	**0.205639**

In Fig 2. Show the graphical representation of the comparison done between different text summarization algorithms with parameters being F-measure, Precision, and Recall using their ROUGE Values.

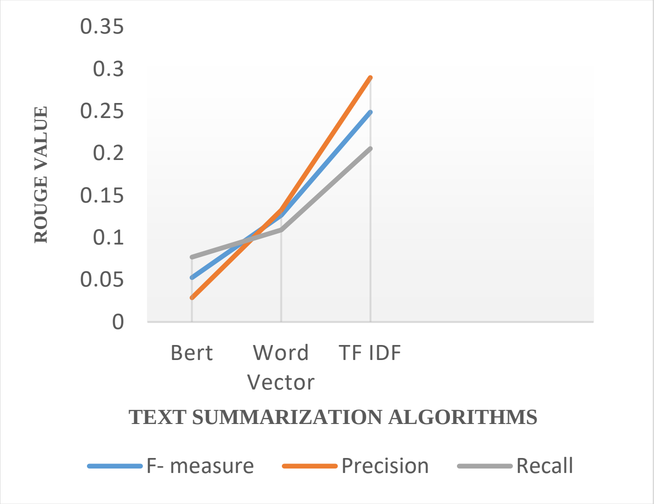

Fig 2. Graphical Comparison

In Fig 3. Shows the user interface of the project and this page consist of different feature which benefits the user while using it the first feature will be the Login and Sign Up buttons which are present in the right top corner. Then comes the Input and output column. User has an option if they want they can paste the complete text in the input column and then get the output or else they can simply upload their files in pdf format of which they want a summary and they just clicking submit the summary will be produced. The Signup and Login feature will be very useful for the user. As the user can sign up and log in to their account and as the user signs in the summaries searched by them are stored in the database and through which they get the access to the past searched summaries and when it comes to storing the summaries only summaries are stored and not any of the credentials of the user.

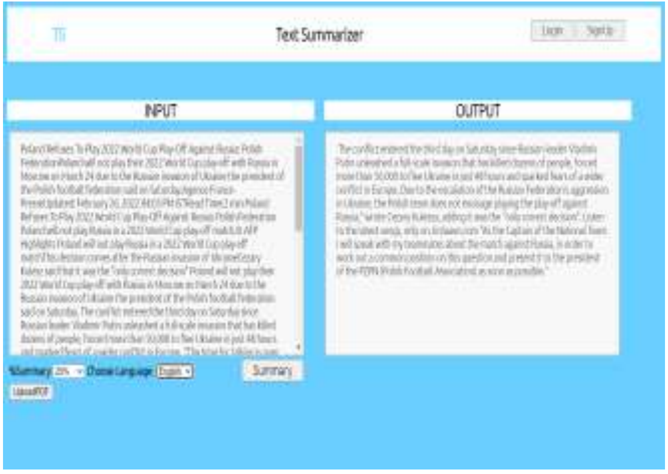

Fig 3. Input and Output GUI

In Fig 3. Shows the output of the summary provided by the user in different languages in this figure the output is in the Hindi language.

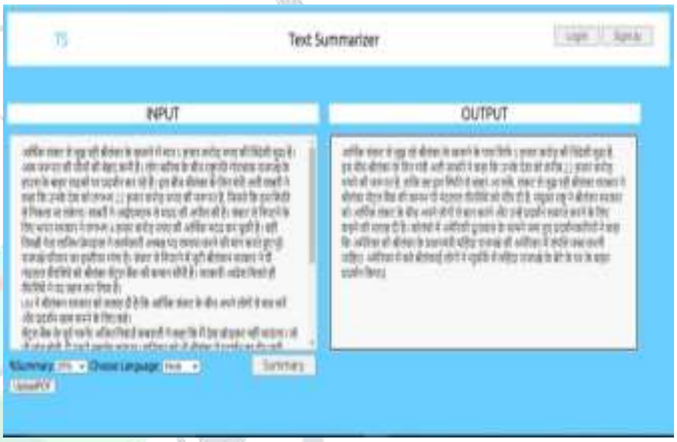

Fig 4. Input and Output GUI (Hindi)

In Fig 5. Shows the history of the user which has been stored by the system and can be accessed by only those users who log in themselves on the system. The user can see the date and time at which the summary was searched and they can also delete any summary which they have searched while being logged in to their account. There is also an admin page where the admin can see all the user's details such as Username, E-mail, Total summaries of the user searched to date, and action to delete the user's details.

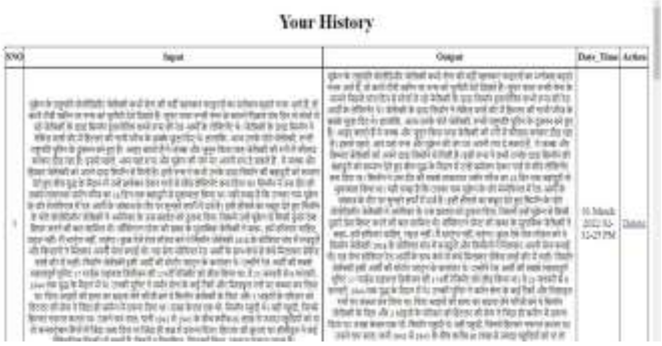

Fig 5. Users History

V. CONCLUSION

On this large informative internet where people used to get the knowledge or inside story of some news. While reading this repetitive information a reader loses their time. Also human summarizing is expensive and may create biased opinions depending upon their thought process. The automatic text summarization is necessary to make a document in condensed form while keeping the main necessary key sentences. So the proper usage of natural language processing to process the information and good techniques to detect the important sentence can lead to producing a summary that is not only short in size but also has meaningful keywords from the original data. Also using the python library for translation will be helpful to make a system that can take more than one language. Thus resulting in the formation of a system that can be used by many people.

VI. REFERENCES

[1] Aditya Jain, Divij Bhatia, Manish K Thakur, (2019), “Extractive Text Summarization using Word Vector Embedding”, *IEEE*

[2]Li-Jyun Chen, Chih-Ying Chen, Hung-Yu Chen, Guan-Yu Chen, Yen-Wen Chen, (2019), "Implementation of Chinese Reader Aid for visually impaired by Using Neural Network and Text Summarization Technologies", *IEEE*

[3]J. N. Madhuri and R. Ganesh Kumar, (2019), "Extractive Text Summarization Using Sentence Ranking", *IEEE*

[4] Sagarika Pattnaik, Ajit Kumar Nayak, (2019), “Summarization of Odia Text Document using Cosine Similarity and Clustering”, *IEEE*

[5]Kasimahanthi Divya, Kambala Sneha, Baisetti Sowmya, G Sankara Rao, (2019), "Text Summarization using Deep Learning", IRJET

[6] A. Jadhav, R. Jain, S. Fernandes and S. Shaikh, (2019), "Text Summarization using Neural Networks", *IEEE.*

[7] Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, and Ayoub Bagheri, (2018), “Query-oriented Text Summarization using Sentence Extraction Technique”, *IEEE*

[8] Meena S M, Ramkumar M P, Asmitha R E and Emil Selvan G SR, (2020), “Text Summarization Using Text Frequency Ranking Sentence Prediction”, *IEEE*

[9] Narendra Andhale and L.A. Bewoor, (2019), “An overview of text summarization techniques”, *IEEE*

[10] Neelima Bhatia and Arunima Jaiswal, (2019), “Automatic text summarization and its Methods”, *IEEE*

[11] Prabhudas Janjanam and CH Pradeep Reddy, (2019), “Text summarization: An essential study”, *IEEE*

[12] Adhika pramita widyassri, Ahmad Zainul fanani, Abdul syukur, Ruri suko basukiEdy, and Noersasongko, (2019), "Literature Review of Automatic text summarization: Research trend, Dataset, and method", *IEEE*