Hybrid Deep Learning Framework for Enhanced Text Summarization

Akshay M J, Swarnagowri M S,Tejaswini N R,Vilasa Bai B G,Supriya H B,Pavan M

Department of Information Science and Engineering

Jawaharlal Nehru New College of Engineering,Shivamogga

***Abstract*—Summarizing technical texts has attracted a lot of research interest, with hybrid methods representing a prominent area. Hybrid summarization is generally divided into two main approaches: extractive summarization and abstractive summarization. Extractive methods focus on selecting and presenting key sentences as-is from the source text, while abstractive summarization involves understanding the context to create a concise, rephrased summary while retaining the main idea of the document. This study investigates the effectiveness of combining these methods using modern machine learning techniques. Advanced models such as TextRank, Kmeans clustering, and BART (Bidirectional Auto-Regressive Transformer) are fine-tuned for summarizing medium articles. We evaluate the performance of these models using ROUGE metrics to assess the accuracy of the generated summaries compared to human-written ones. Additionally, we investigate the outcomes of combining these models in different configurations to assess the benefits of a hybrid strategy for summarization.**

***Key words*-Deep learning,Extractive summarisation,**

***Abstractive summarisation,K*-means clustering,TextRank**

I.INTRODUCTION

This Text summarisation has emerged as a crucial area of research within natural language processing (NLP), owing to the ever-increasing volume of textual data generated daily. Summarisation techniques aim to condense large bodies of text into shorter, coherent, and information-rich summaries while retaining the essential meaning.Traditional approaches to summarisation fall into two main categories: extractive and abstractive. Extractive summarisation involves selecting key sentences or phrases directly from the source text, whereas abstractive sunumarisation generates summaries using paraphrased or newly constructed sentences. However, these methods have limitations when applied independently. prompting the need for hybrid approaches that combine the strengths of both techniques [12].

The advent of deep learning has revolutionised text summarisation by enabling models to understand and generate human-like text. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks were initially employed for sequence-to-sequence tasks, including summarisation [6]. Subsequently, adv advancements in transformer-based architectures, such as BERT
(Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have further enhanced the capabilities of deep learning models [18] These models excel at capturing contextual relationships within test. making them highly effective for abstractive sunmarisation tasks. However, purely abstractive approaches often face challenges in maintaining factual accuracy, especially for complex documents [7]. Hybrid text summarisation seeks to address these challenges by leveraging the complementary strengths of extractive and abstractive methods. For example, an influential extractive summarisation technique that works by constructing a graph of sentences, where each node represents a sentence, and the edges between nodes reflect the similarity between sentences. It has been widely used for its winery computational efficiency and simplicity in producing extractive summaries.

II.RELATED WORKS

[1] The research focuses on various aspects of automatic text summarization (ATS) and its different classifications. It highlights the importance of summarization. The document mentions that single document text summarization is easier to implement compared to multi-document summarization, which is a complex task. It also introduces the concepts of abstractive and extractive summarization methods, along with hybrid methods that combine both approaches. In terms of summarization methods, the document explains that extractive summarization involves selecting important sentences from the source text, while abstractive summarization involves paraphrasing and generating new sentences.

[2] The work introduces an innovative approach called ATSDL for Abstractive Text Summarization (ATS). ATS involves creating concise summaries by merging information from various source sentences. ATSDL, based on LSTM-CNN, stands out by exploring fine-grained semantic phrases in two stages.Firstly, it extracts phrases from source sentences; secondly, it generates text summaries using deep learning. Experimental results on CNN and Daily Mail datasets demonstrate that ATSDL outperforms existing models in both semantics and syntactic structure, achieving competitive results in manual linguistic quality evaluation.

[3] The study inveastigates the evaluation of ASDKGA is conducted on two distinct corpora, namely KALIMAT and Essex Arabic Summaries Corpus (EASC). The assessment employs the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework, comparing automatically generated summaries by ASDKGA with human-generated summaries.Additionally, the ASDKGA

approach is benchmarked against three other existing Arabic text summarization methods.

Results showcase the effectiveness of ASDKGA in summarizing Arabic political documents, with an average Fmeasure of 0.605 achieved at a compression ratio of 40%.

[4] This work discusses the topic of text summarization, which involves extracting important information from a document or set of documents and presenting it in a concise form. The document highlights different approaches and techniques used in text summarization, such as clustering algorithms, graph-based ranking, and unsupervised learning methods. It emphasizes the importance of covering all topics in the text, avoiding redundancy, and providing diversity in a summary. It concludes by discussing the proposed COSUM model, which is a two-stage sentence selection model based on clustering and optimization techniques.

[5] This research focuses on the challenges and approaches to text summarization using deep learning models. It highlights the limitations of traditional methods. These methods also lack semantic information compared to distributed representations of words and sentences. However, deep learning models come with their own shortcomings, including the requirement of a large amount of computational resources and the need for labelled training data. The document proposes a document summarization framework that does not require model training.

[6] This research explores Abstractive text summarization is the task of creating a summary from a document by merging facts from different sources and make a short description of them. In this procedure, the meaning and the content information should be kept. In this paper, a hybrid summarization system using deep recurrent neural network is proposed, which can create new sentences by information extracted from the text. The proposed model is the combination of extractive and abstractive summarization and has the encoder-decoder structure. The encoder extracts information from the source document and encodes this information in a compressed representation. The decoder takes the encoder ' s output as input and generates a summary, which has an acceptable semantic and syntactic structure.

[7] Thework explores automatic summarization of technical articles is a field that has garnered a fair amount of interest, and one that enjoys a significant portion of NLP-related research. As a whole, automatic summarization can be split into two broad categories - extractive and abstractive. Extractive summarization implies that important and relevant sentences are picked from the article as is, and inserted in the summary. Abstractive summarization, on the other hand, requires contextual understanding of the document, and rearranging and shortening the sentences, while maintaining the core essence of the article. Multiple algorithms have been proposed for both these classes of automatic summarization. In the recent past, the emergence of pretrained language models for NLP tasks have been heralded by the creation of attention mechanisms and Transformers.

[8] This research explores On-line information has increased tremendously in today's age of Internet. As a result, the need has arose to extract relevant content from the plethora of available information. Researchers are widely using automatic text summarization techniques for extracting useful and relevant information from voluminous available information, it also enables users to obtain valuable knowledge in a limited period of time with minimal effort. The summary obtained from the automatic text summarization often faces the issues of diversity and information coverage. Promising results are obtained for automatic text summarization by the introduction of new techniques based on graph ranking of sentences, clustering, and optimization.

[9] This work focuses sequence-to-sequence (seq2seq) models have gained a lot of popularity and provide stateof-the-art performance in a wide variety of tasks, such as machine translation, headline generation, text summarization, speechto-text conversion, and image caption generation. The underlying framework for all these models is usually a deep neural network comprising an encoder and a decoder.

[10] The work presents Natural Language Processing is vast area which has great importa importance when people started to interpret human language from one form another. Summarization is one of the research works in NLP which concentrates on providing meaningful summary using various NLP tools and techniques. Since huge amount of information is used across the digital world, it is highly essential to have automatic summarization techniques. Extractive and Abstractive summarization are the two summarization techniques available. A lot of research works are being carried out in this especially in extractive summarization.

[11] The work addresses currently used metrics for assessing summa rization algorithms do not account for whether summaries are factually consistent with source documents. We propose a weakly-supervised, model-based approach for verifying factual consistency and identifying conflicts between. source documents and a generated summary. Training data is generated by applying a series of rule-based transformations to the sentences of source documents.

[12] This study explores summarization, is to reduce the size of the document while preserving the meaning, is one of the most researched areas among the Natural Language Processing (NLP) community. Summarization techniques. the basis of whether the exact sentences are considered as they appear in the original text or new sentences are generated using natural language processing techniques, are categorized into extractive and abstractive techniques.

[13] This study addresses text summarization is a subtask of natural language processing referring to the automatic creation of a concise and fluent summary that captures the main ideas and topics from one or multiple documents. Earlier literature surveys focus on extractive

approaches, which rank the top-n most important sentences in the input document and then combine them to form a summary.

[14] This research focuses text summarization plays an important role in the area of natural language processing. The need for information all over the world to solve specific problems keeps on increasing daily. This poses a greater challenge as data stored on the internet has gradually increased exponentially over time. Finding out the relevant data and manually summarizing it in a short time is a challenging and tedious task for a human being. Text Summarization aims to compress the source text into a more concise form while preserving its overall meaning.

[15] This research investigate cloud resources, such as webpages, blogs, news, user messages, and social network platform, have accumulated gigantic amounts of textual data, and they are increasing exponentially every day. In addition, various articles, books, novels, legal documents, scientific papers, biomedical documents, and other archives also contain rich textual content. As a result, information overload is becoming more and more serious.

[16] This research investigate a class of neural networks known as Recurrent Neural Networks (RNNs) are capable of processing sequential input, including time series and plain language. In the shortest possible time to find relevant and useful information, it is for sure very helpful if the information is. summarized, but it typically requires a lot of effort, dedication, patience, and attention to detail for humans to go through and summarize the lengthy texts.

III.PROPOSED APPROACH

A.System Architecture

The system architecture for the hybrid text summarisation model combines two come modules extractive and abstractive summarisation.The extractive module identifies the most important sentences from the input document, while the abstractive module generates a concise summary by rephrasing the most relevant content. The integration of these two modules ensures that the model generates summaries that are both fluent and contextually relevant. The hybrid system follows a pipeline approach. where the input text is first processed by the extractive summarisation module to identify important content[11]. The extractive summaries are then passed to the abstractive module, which refines the content to generate a coherent, fluent summary. This approach benefits from the strengths of both extractive and abstractive techniques, ensuring that the final output is both factual and redable.

B.Data Set selection and Preprocessing

For training the summarisation model, we use a wellestablished dataset suited for both extractive and abstractive summarisation tasks. One such dataset is the CNN/Daily Mall dataset, which contains news articles paired with human-written summaries. This dataset is widely used for summarisation tasks due to its size and diversity, covering a wide range of topics and writing styles[16].

Data cleaning

Data cleaning is a critical preprocessing step to ensure that the input text is an optimal form for training deep learning models. The raw text data often contains irrelevant content, such as advertisements, metadata, and noisy text. Therefore, the cleaning process involves removing noncontent elements like HTML tags, URLs, and special characters. Additionally, any duplicate data or poorly formatted content is discarded to improve the improve the quality of training data[3].

Tokenisation

Once the data is cleaned, tokenisation is performed to convert the text into a format that can be processed by deep leaming models, Tokenisation involves splitting the text into individual words or subwords, which can then be represented as numerical vectors[9]. This process is essential for enabling the model to understand and process the textual data efficiently. WordPiece of Byte Pair Encoding (BPE) tokenisation techniques are often employed to deal with outof-vocabulary words by splitting them into subword units.

C.Hybrid model design

The design of the hybrid summarisation model involves the combination of two modules extractive summarisation and abstractive summarisation. Each module is trained separately and integrated to produce a final summary.

Extractive Summarization Module

The extractive summarisation module selects key sentences or phrases freen the input document 1. It relies on sentence embeddings or TF-IDF values to identify important sentences that contribute to the meaning of the document BERT or RoBERT can be fine-tuned for extractive summarisation tasks allowing the the model to learn the most salient sentences based on contextual relationships[15].

Abstractive Summarization Module

The abstractive summarisation module generates a concise summary by rephrasing or paraphrasing the extracted content. This module uses sequence-to-sequence(seq2seq) models, often enhanced with an attention mechanism ensure that the generated summary remains coherent and contextually relevant. Models such as GPT orT5 commonly used for abstractive summarisation tasks, where the model is trained to produce fluent summaries by capturing the underlying meaning of the input text[13].

Integration of Modules

To combine the two approaches, the output extractive summarisation module (the exacted se passed as input to the abstractive summarisation.This hybrid approach allows for accurate and content extraction, while the abstractive module the summary is fluently written and contextually , integration of these two modules is key to improving the quality of the summary[2].

IV.IMPLEMENTATION

The input text is pre-processed (tokenization, stop-word elimination, and stemming) before the procedure starts. Key statements are identified by extractive summarization and rephrased by abstractive summarization. Sentences are scored using the Text Rank algorithm, and redundancy is decreased using K-means clustering. The final golden summary is the result of the BART model's refinement of the summary. To guarantee quality, the output is assessed using ROUGE metrics[10].

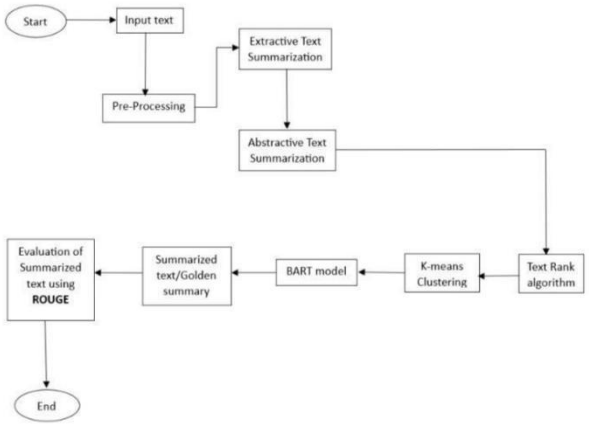

Fig 1. Implementation of Hybrid text summarixation

The above Figure1 shows the system design of hybrid text summarization using deeplearning. Text summarization includes input text from user, pre-process the text, and summarize the text using text summarization methods. These are explained below in detail

Using TextRank, the K-means method for extractive summarization, and a pre-trained model for abstractive summarization called BART (Bidirectional AutoRegressiveTransformer), the "Hybrid Approach for Text Summarization using Deep Learning" has been put into practice. Below is an explanation of the above mentioned models and algorithms.

A.TextRank : TextRank is an unsupervised, extractive method for summarizing texts. This method would separate the text into discrete sentences after first concatenating all of the content found in the articles. The next stage will be to identify each sentence's vector representation, or word embeddings. Vectors are then computed and saved in a matrix by examining text similarities. For the purpose of calculating sentence rank, the similar matrix is then transformed into a graph, with sentences serving as vertices and similarity scores as edges.Ultimately, the final summary or paragraphs are composed of a specific amount of highly scored sentences[14].

Clustering using K-Means :One of the most popular and straightforward clustering methods is K-means. This kind of partitioning clustering technique divides the provided text or summary into arbitrary sections. A more reliable and quick approach for creating spherical clusters is K-means. The number of clusters must be entered at the start, however it is optional if it is adjusted[13]. Each data point is iteratively assigned to the closest cluster centroid by the algorithm, which then recalculates the centroids using the freshly created clusters. And for the extractive summarization technique, those freshly created clusters are referred to as the final output or final summary.

The BART : Both a left-to-right decoder (like GPT) and a bidirectional encoder (like BERT) are part of the sequencetosequence/machine translation architecture of BART. It is a denoising auto encoder for pre-training sequence-to- sequence models. It is learned by using a noising function to corrupt text and then learning a model to reconstruct the original text. Using a novel in-filling method that substitutes a single mask token for text spans, the pretraining job involves randomly rearranging the source words. Transformer neural machine translation is the foundation upon which it is based[[4]. In addition to its notable use in text production, BART performs well on comprehension tasks.

V.RESULTS AND DISCUSSION

The below table 1 shows the input and output number of words.

Table 1 Input and Output count of words

Technique	Input	Output
Hybrid	49001	1236

ROUGE Scores for table 1 has been demonstrated in the below table 2. Where it has ROUGE-1, ROUGE-2 and ROUGE-L as columns for that table. The above table contain values which are said to be count of words in the both reference summary and generated summary.

Table 2 ROUGE Scores for table 1

Technique	ROUGE-1	ROUGE-2	ROUGE-L
Hybrid	0.0189	0.0179	0.0185

The above ROUGE scores for hybrid text summarization. The input text contains 49,001 words, while the reference text has 1,236 words. The results vary depending on the input's word count. Table 3 presents updated ROUGE scores after reducing the word count of the input, compared to the results shown in Table 1.

Table 3 Input and Output count of words

Technique	Input	Output
Hybrid	13480	3653

ROUGE Scores for table 3 has been demonstrated in the below table 4. Where it has ROUGE-1, ROUGE-2 and ROUGE-L as columns for that table. The above table contain values which are said to be count of words in the both reference summary and generated summary.

Table 4 ROUGE Scores for table 3

Technique	ROUGE 1	ROUGE 2	ROUGE L
Hybrid	0.1146	0.1487	0.1308

The above ROUGE scores are performed on the text, which has 13480 words and output for the reference text is 1044 words. The result is obtained after applying both the techniques on the reference text. Remember, the output of the hybrid text summarization may vary based on the count of words in the input or reference text.

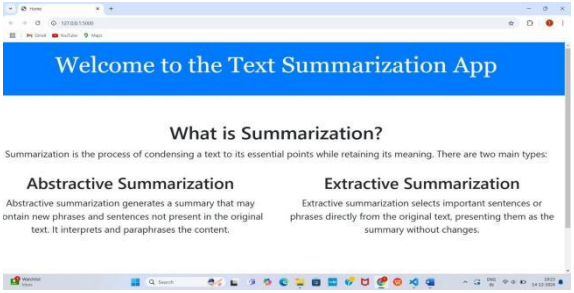

Fig 2 Home page

Figure2 shows that the home page welcomes users to a Text Summarization App It explains summarization as making a text shorter while keeping the main meaning

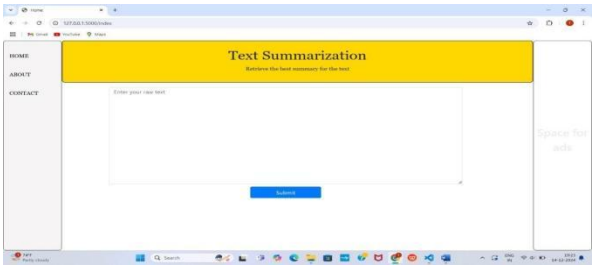

Fig 3 Index page

Figure3 shows that this page is the Text Summarization tool where users can enter their text to get a summary there is a text box to type or paste your raw text.

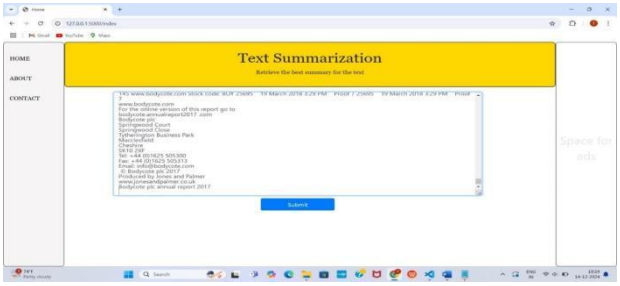

Fig 4 Index Text

Figurer 4 shows that the web page, where user has provided the text for summarization. The provided summary was ready for the summarization once user click the submit button.

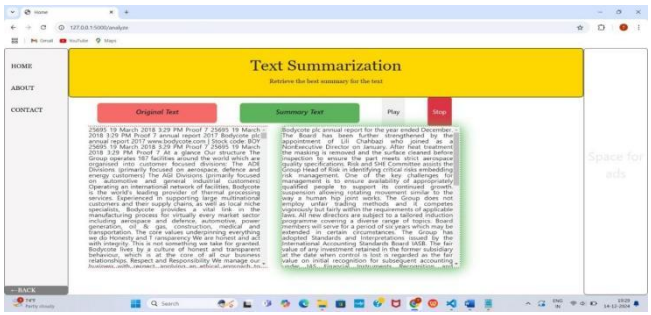

Fig 5 Output Page

Figure 5 shows the output page, where it contains the both generated summary and the reference summary side by side.

VI.CONCLUSION AND FURTHER STUDY

This study has covered a wide range of subjects related to extractive and abstractive article summarization. Pre-trained language models are examples of contemporary techniques that have been examined and put into practice. Individual models have been altered in an attempt to achieve the greatest outcomes, and combinations of the top-performing models have also been used[1].

As indicated in the previous section, the investigation's findings show that our algorithms' success levels with extractive summarization differ. Their distinct performances show that they are more appropriate for various task requirements[9]. They also demonstrate that their capacity is greatly influenced by the initial dataset and pre-training methodology. The hybrid model's use of extractive and abstractive techniques demonstrates that the combined results are significantly superior to their separate efforts. This suggests a technique that could lead to better outcomes on many NLP tasks and an area that needs more study.

The performance of pre-trained language models, conventional techniques, and machine evaluation of outcomes are only a few of the subjects that have been examined in this research on text summarization. It raises a number of fresh issues and suggests fresh areas for research, such as the effectiveness of quantitative assessment techniques and the advantages of a hybrid approach. An effective method for applying machine learning novelty to essentially NLP problems is to leverage pre-trained language models. Nonetheless, a number of adjustments can be done within the same field to improve the result[5]s. To get the greatest summary with just one language model, these changes can be made to individual models. Another option is to try combining several models and evaluating how well they work together. Recent innovations for the task of text summarization include graph-based algorithms, deep learning approaches, and combining extractive and abstractive techniques for optimal output summaries. Although in their state of research, all of these practices promise greater efficiency and more precise results in the eventual future.

VII.REFERENCES

[1] Mansoora majeed, Kala MT, "Comparative study on extractive summarization using sentence raking algorithm and text ranking algorithm", IEEE, 2023.

[2] Rohan habu, Rohit ratnaparkhi, Anjali askhedkar, Sunitha Kulkarni, "A hybrid extractive-abstractive framework with pre-post-processing techniques to enhance text summarization" IEEE, 2023.

[3] S. Song, H. Huang, and T. Ruan, ''Abstractive text summarization using LSTM-CNN based deelearning'' Multimedia Tools Appl, vol. 78, no. 1, pp. 857–875, Springer, Jan. 2019.

[4] Al-Radaideh, Q. A., & Bataineh, D. Q, "A Hybrid Approach for Arabic Text Summarization Using Domain Knowledge and Genetic Algorithms". Cognitive Computation, Springer ,10(4), 651–669, 2018.https://doi.org/10.1007/s12559-018-9547-z

[5] Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N., "COSUM: Text summarization based on clustering and opt optimization. Expert Systems", Article, 36(1), 1–17. 2016. https://doi.org/10.1111/exsy.12340

[6] Myeongjun Jang and Pilsung Kang, "Learning-Free Unsupervised Extractive Model", IEEE, Jan 13, 2021.

[7] Majid Abolghasemi, Chitra Dadkhah and Nasim Tohidi, "HTS-DL: Hybrid Text Summarization System using Deep Learning", IEEE, 2022.

[8] Jigisha M Narrain, Vanshika Taneja, Sanjana B Atrey, Jahnavi Sivaram, Dinesh Singh, "Extractive Summarization - A Comparison of Pre-Trained Language Models and Proposing a Hybrid Approach", IEEE,2023.

[9] Mengli Zhang, Gang Zhou, wanting Yu, Ningbo Huang, Wenfen Liu, "A Comprehensive Survey of Abstractive Text Summarization Based on Deep Learning", Computational Intelligence and Neuroscience, Volume 2022.

[10] Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, "Evaluating the Factual Consistency of Abstractive Text Summarization", Salesforce Research ,2019.

[11] Nikolaos Giarelis, Charalampos Mastrokostas, Nikos Karacapilidis, MDPI, Basel, Switzerland, 2019.

[12] Som Gupta, S. K Gupta, "Abstractive Summarization: An Overview of the State of the Art", 2019.

[13] G. Karuna, M. Akshith, Parige Sai Dinesh, Bodhan Vishnu Vardhan, Yashwant Singh Bisht, M. N. Narsaiah, "Automated Abstractive Text Summarization using Deep Learning" Web of Conferences, Elsevier, 2023.

[14] Mohdkhizir siddiqui, amreen ahmad, om pal, tanvir ahmad, "CoRank: A clustering cum graph ranking approach for extractive summarization", Association for Computing Machinery, 2021.

[15] Yaser Keneshloo, TianShi, Naren Ramakrishnan, and Chandan K. Reddy, "Deep Reinforcement Learning for Sequence-to-Sequence Models", 2019.

[16] James Mugi Karanja and Abraham Matheka, "A Hybrid Model for Text Summarization Using Natural Language Processing", Center for Open Access in Science – Open Journal for Information Technology, 2022.