**Research Proposal/Synopsis for MS Thesis**

**Department of Computer Sciences**

**1. Name of the Student:** NABILA ANUM

**2. VU ID:** ms220400077 **3. Session:** 2022-2024

**4. Semester:** 4th **5. Field of Specialization:** Databases

**6. Title of Research Proposal:** “Automatic Text Summarization of Educational Books by Fine-Tuning of Abstractive Models”

**7. Date of Enrolment in Research:** 12.11.2023

**8. Duration of Proposed Research:** 8 months

**9. Total Funds Requested (if any) Rs. \_\_\_\_\_\_\_NIL\_\_\_\_\_\_\_\_ (**Rupees **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)**

**===============================================================================**

**Supervisor, Supervisory Committee (SC) Information**

**1. Name of Supervisor:** Mushtaq Hussain

**Designation:** Assistant professor **Email ID:** mushtaq.hussain@vu.edu.pk

**Affiliation:** Virtual University of Pakistan

**2. Name of Supervisory Committee (SC) Member 1:** Dr Said Nabi

**Designation:** Lecturer **Email ID:** said.nabi@vu.edu.pk **Affiliation:** Virtual University of Pakistan

**3. Name of Supervisory Committee (SC) Member 2:** Mehboob Ali

**Designation:** Lecturer **Email ID:** mehboob.ali@vu.edu.pk

**Affiliation:** Virtual University of Pakistan

**Topic**

“Automatic Text Summarization of Educational Books by Fine-Tuning of Abstractive Models”

**Abstract/Summary**

Several challenges are associated with text summarization systems that have already been designed by using Natural Language Processing techniques and algorithms in the past. An efficient model must be able to understand where important information exists in the given text to produce an innovative summary. A supervised learning model to enhance the efficiency of text summarization out of the text will be proposed in this research. The experiments involving a combination of extractive and abstractive approaches that lead to hybrid models with various sizes of input to evaluate the model. The main purpose of this research is to fine-tune the T5 and Bart models that are used for text summarization. The dataset to be used for the implementation of hybrid models will be the topics from educational books. The results will be evaluated through the well-known ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric. The expected results would be more than 0.69 to 0.70, as F score calculated by ROUGE. The significance of this study is that it can turn large amounts of data into easy-to-use summary information, and it can aid students at higher institutions in comprehending complex concepts by providing concise and accessible explanations.  Similarly, teachers can take advantage of summarized material by making notes from different sources for students or taking important highlights, which will be helpful for students.

**Introduction**

Text summarization leads to the process of searching for important information from the source document and then converting this information into a precise format without updating the importance of the original text, as stated by Zaware et al., (2021). Similarly, Elbarougy et al., (2020) also stated in their paper that summarization of textual data is the process of reducing the number of texts to retrieve the important information from the input text and presenting them to users. Automatic text summaries run summaries automatically. Hence, summarization plays a vital role and is the most efficient solution in the case of literature review, for getting the gist of information covered in newspapers, covering the maximum content from story books as mentioned by Zhao et al., (2022) or from course content materials for prior exam preparation. Similarly, while reading a newspaper, it is ultimately necessary to get the gist of important news in the form of outlines. Furthermore, while giving some useful information to students from books available in the school library, it is sometimes necessary to cover all important details in a limited time duration. Moreover, students at the college or university level may need to consult multiple books for a single course, and everyone wants to cover a maximum of the respective books for exam preparation.

In the field of education, this has led to the rise of e-learning in many higher education institutions, as students are encouraged to access research and learning materials online rather than visiting physical libraries. This can lead to more opportunities for research and learning when students use a variety of online learning tools. This requires a solution that can help transform large amounts of data into user-friendly data summaries, as suggested by Karanja and Matheka (2022). Likewise, the materials on the internet change rapidly, and many repetitions or inconsistencies are encountered in this information. Hence, the need for summarizing and organizing written content becomes pressing and significant. Manual summarization proves to be costly and time-intensive. Text summarization emerges as a pivotal solution to address this challenge, as highlighted by El-Kassas et al., (2021). Moreover, automated text summarization would be helpful for a lot of tasks. Machine Learning has introduced numerous techniques and models for text summarization in Natural Language Processing. Basically, machine summarization is categorized into two groups, i.e. Extractive summarization and abstractive summarization, as mentioned by Gopikakrishna et al., (2021). An extractive approach to text summarization involves the selection of sentences and phrases from the original text to form a summary. It has multiple phases before it generates a summary that include preprocessing (removal of stop words, stemming, and tokenization), calculating scores for sentences, and then, last but not the least, is the generation of important sentences based on the calculated scores of these sentences. While the abstractive approach includes joining sentences from original text and combining it with newly generated sentences to form more innovative summary. Unlike extractive methods, which select and combine existing sentences, abstractive methods aim to produce summaries that may not exist word-for-word in the original text. Abstractive summarization often employs encoder-decoder architectures, such as sequence-to-sequence (Seq2Seq) models, which consist of two main components: encoder and decoder. To implement an abstractive approach, the model is trained using suitable parameters, and then it is evaluated using Rouge (Recall-Oriented Understudy for Gisting Evaluation).

As far as this research is concerned, text summarization will be performed using an extractive abstractive approach. This involves using a technique called a Word Frequency Algorithm to extract information. A Word Frequency Algorithm is used to retrieve important sentences by calculating their scores, and then these retrieved sentences are passed through a seq-to-seq transformer that is fine-tuned to improve its accuracy. In this research the accuracy of the T5 and BART models will be improved using a novel technique called a hybrid approach. Most of the literature on text summarization includes either the extractive or the abstractive approach. This research includes the summarization of the topics from educational books, so it will be really helpful for the students of higher institutions to cover important contents from multiple books for the preparation of exams. Similarly, by utilizing the seq-to-seq transformer model to generate summaries of topics from a textbook, the thesis contributes to educational enhancement. Summaries can aid students at higher institutions in comprehending complex concepts by providing concise and accessible explanations, thereby improving learning outcomes. Summaries generated by some seq-to-seq models can guide educators in prioritizing content and designing instructional materials.

**Research Question(s)**

Q#1: How to implement a hybrid method for text summarization that combines the Word Frequency Algorithm with transformer models like T5 and Bart?

Q#2: How to enhance the accuracy of abstractive models T5 and Bart by fine-tuning on self-created dataset?

**Research Objectives**

The purpose of this research is to produce a summary of the input text that contains important sentences and to display important highlights from the given text. To include important sentences in summary firstly, extractive approach will be used i.e. Word frequency Algorithm and then after retrieving the important sentences using this approach, the abstractive approach i.e. seq-to-seq transformer models that are T5 and Bart will be used to make the summary more innovative and precise.

**Review of Literature**

The review of the literature concerning the text summarization shows that the work done in the context of summarization is performed either by using an extractive or an abstractive approach. Likewise, Zhao et al., (2022) have presented a novel framework in which questions from story books are generated through prior summarization. The described framework contains three modules: question type distribution learning, event-centric summary generation, and educational question generation. The approach used in this paper is fine-tuning of BART for two times. First time for Event-Centric Summarization and second time for automatic generation of questions out of the summarized text. The performance of the proposed mechanism is evaluated through the comparisons between generated questions from proposed model and the questions from Fairy tale QA dataset. The calculated ROUGE L on F1 is 41.78/38.29.

Similarly, some details about previously used models, for example, RNN, CNN, etc., were mentioned by Etemad et al., (2021). But due to drawbacks such as gradient vanishing, exploding, long term dependencies, parallelization, and being computationally constrained, these are replaced by recent transformers, for example, BERT and other seq- to-seq and text-to-text transformers.In this paper, the T5 model is trained on new datasets that are XSUM and GIGAWORD.Xsum dataset contains 204045 articles for the training set, 11333 articles for testing, and the same number of articles for the validation set. In this experiment, the model is trained and tested on the whole dataset. Gigaword is a collection of English articles with summaries that contains 3 million articles. In this experiment, T5 is fine-tuned and the accuracy of the fine-tuned model is measured by ROUGE1, 2, L and ROUGE LSUM. The maximum of these scores is 43.02.

Furthermore, Deokar & Shah (2021) have compared two abstractive transformer models that are t5 and Bert. They have made use of the BBC News Summary dataset, also available on Kaggle. This dataset has news articles grouped into 5 classes – ‘business’, ‘entertainment’, ‘politics’, ‘sport’, and ‘tech’. There are also summaries provided for each article in each class. The authors have mentioned in newspaper that they have also introduced a method of scrapping news articles. The accuracy of BART and T5 is compared using ROUGE F1 scores that are 33 % and 26% respectively.

Similarly, Karanja & Matheka (2022) have described two approaches to text summarization that are extractive and abstractive. To have good precision and recall, the authors have implemented a hybrid approach for text summarization. This hybrid approach means that the test is first summarized through an extractive approach, and then the output of this will act as the input for abstractive summarization. The improved scores of precision and recall are shown in tabular and graphical form. And the accuracy of their proposed approach is given as F1 SCORE: ROUGE 1=0.232166667, ROUGE2=0.048533333, ROUGE=0.1416.

Moreover, Guoy et al., (2022) have presented a new Transformer-based neural model called LongT5, with which the effects of scaling both input length and model size can be explored at the same time. Through experimentation in various complex summarization and question answering datasets, the authors have explored the performance gains that can be achieved by scaling both input length and model size, resulting in state-of-the-art results on multiple datasets. The results to show the accuracy are computed on arXiv, PubMed, BigPatent, MediaSum, and TriviaQA. The main conclusion on the result is LongT5 (xl - 16k input) =48.35, 21.92 and 44.27.

In the same vein, the authors Mengi & Kakade (2023) have used a pre-trained model proposed by Colin Raffel that is fine-tuned in the Xsum and Gigaword datasets and produces state-of-the-art performance and abstractive text summarization. Xsum is a collection of BBC articles, with a summary written by the author of the article. The dataset consists of a huge collection of BBC News articles, Amazon product reviews, IMDB movie reviews, and English sentiment data, amounting to a significant 2.5GB in size and spanning several million rows. The data set is first prepared for training, passed through multiple stages of preprocessing, including cleaning, stemmatization, and lemmatization, and appointed title. The researchers have tried to achieve an improvement in the summarization model by increasing ROUGE from 13.6168 to 21.4997. Conversely, the RoBERT a model excelled in sentiment analysis, achieving a remarkable accuracy of 94.71% and an F1 score of 94.60%.

Additionally, the use of the T5 model is observed by Kumar et al., (2022). It first needs a fine-tuning of the t5 model on a small dataset, and then the fine-tuning of T5-eQA is needed to generate the precise questions. The accuracy is computed by BLEU and ROUGE, which are given by T5QG =40.96, 17.54, 19.21, 42.36, and Contrastive\_T5QG = 42.04, 19.11, 20.07, and 48.50.

Similarly, to bring light on text summarization Kaur & Sharma (2023), have proposed a novel framework called SA model implementing deep learning for CRS (SADL-CRS) based on the effective approach of SA and review summarization (RS). There are three stages mentioned in this paper: the input stage, the tokenization stage and the last one is the summarization section. In the Input section, data is pre-processed, whereas in the tokenization section, the pre-processed data is tokenized and marked in the review as positive, negative or neutral. Moreover, in the last summarization section, the final summary of the given review is retrieved. The performance is evaluated on ROUGE score with respect to the multiple datasets, for example, the SemEval-2014 dataset and the Sentiment140 dataset. And the accuracy is determined as ROUGE L OF F1 CRS 18.71.

Additionally, to achieve efficiency in text summarization Liu et al., (2021) have presented a novel framework. The proposed model, Refactor, can be used either as a base system or a meta system, effectively coping up with the learning gaps introduced in the two-stage learning. The mainly used techniques BART, GSum and PEGASUS are used as the base systems keeping in view their capability of achieving improved performance on at least one dataset. The datasets which are specifically used in this research paper are Xsum, Pubmed and Wikihow.And the measure used to evaluate the performance is used is Rouge. The maximum accuracy is achieved on Pubmed dataset by fine tuning the model is 43.72.

Moreover, the analysis revealed several unexpected trends in the text summarization. Rehman et al., (2023) have presented the analysis of three pretrained models named as google/pegasus-cnn-dailymail, T5-base, facebook/bart-largecnn on CNN-dailymail on three different datasets that are Samsum, Billsum and CNN-dailymail.The highest accuracy in the form of f1 score is achieved by BART model on Billsum dataset that is 42.84.

Furthermore, the authors Merrouni et al., (2023) have focused on the challenges that persist in the development phases of text summarization. These include the detection of text relevancy, identification of redundancy and coherency, and the third challenge pertains to abstractive and hybrid summarization. In this paper, they introduce EXABSUM, an ATS SYSTEM equipped to generate two distinct summary categories that follow extractive and abstractive approach respectively. EXABSUM Extractive includes multiple preprocessing phases such as stemmatization, lemmatization, calculation of TF/IDF for each term and calculation of relative position of sentences in the summary. Similarly, EXABSUM Abstractive includes sub stages that are Word graph generation, paths filtering, re-ranking using key phrases and then finally the abstractive summary generation. The maximum calculated Rouge score using scoring techniques ISF, sentence’s length, sentence’s resemblance to the title is 0.493 on the DUC2002 dataset.

Text summarization is an interesting research topic among the NLP community that helps produce concise information. For new research to conduct, it is very important to have advance knowledge of the topic on which research is to be conducted along with the review of previous work done on that topic. For this purpose, Widyassaria et al., (2022) have presented the latest research and progress in the field of text summarization through Systematic Literature Review (SLR) method. With SLR method it is proven that it provides the up-to-date and well-equipped knowledge about the use of Text summarization techniques, datasets, and preprocessing techniques feature extraction techniques and the problems and methods to evaluation that are used in previous research regarding text summarization. It also paves a path towards future research.

To highlight the importance of text summarization in the English and other text mining courses of computer science, Abidin et al., (2022) have implemented extractive approach that is TF-IDF with the achievement of a recall value of 0.675, precision value of 1.00 and f-measure test value of 0.8059 which indicates that the summarizing system has vital importance and can be used as a medium for learning English courses and text mining. Summary of a given text is generated by passing from multiple sequential stages that are counting the words in the text, removal of stop words, stemming process, calculation of terms and document frequencies and then formulation of summary using these frequencies. Evaluation of the used technique is performed by making a comparison between system generated summary and online summary by some online tool for example tool4noob.

Text summarization can be possibly achieved through extractive approaches. One of these approaches has mentioned by Christian et al., (2016). In his paper the authors have implemented TF-IDF method for summary generation. The generated summary by the program is compared with the human generated summaries and the online tools. The accuracy achieved is calculated as a F1-score that is 67%.

Moreover, financial documents and annual reports may contain some useful but lengthy information that needs to be summarized in an efficient way. To tackle this problem, an approach is developed by Litvak et al., (2020). The research aims to address the challenge of efficiently summarizing these documents by utilizing TF-IDF weighting of both single-word and multi-word expressions to identify key terms and sequences that capture the most important information in the documents. The F score is used to measure the accuracy of newly developed approach that is 0.433.

Similarly, to generate high-quality and information-rich summaries of numerically-oriented commodities reports the transformer models of Natural Language Processing that are Bart and T5 have fine-tuned by Zeng (2024). The goal is to assess how effectively these models can relay numerical information in summaries, with a focus on aiding farmers, producers, and small businesses in the agricultural industry. The accuracy achieved is Flan T5 vs Numerical=56% Flan T5 vs compression=62% and Bert vs Numerical=17.25% Bart vs compression=22%.

**Table 1: Table of literature Review**

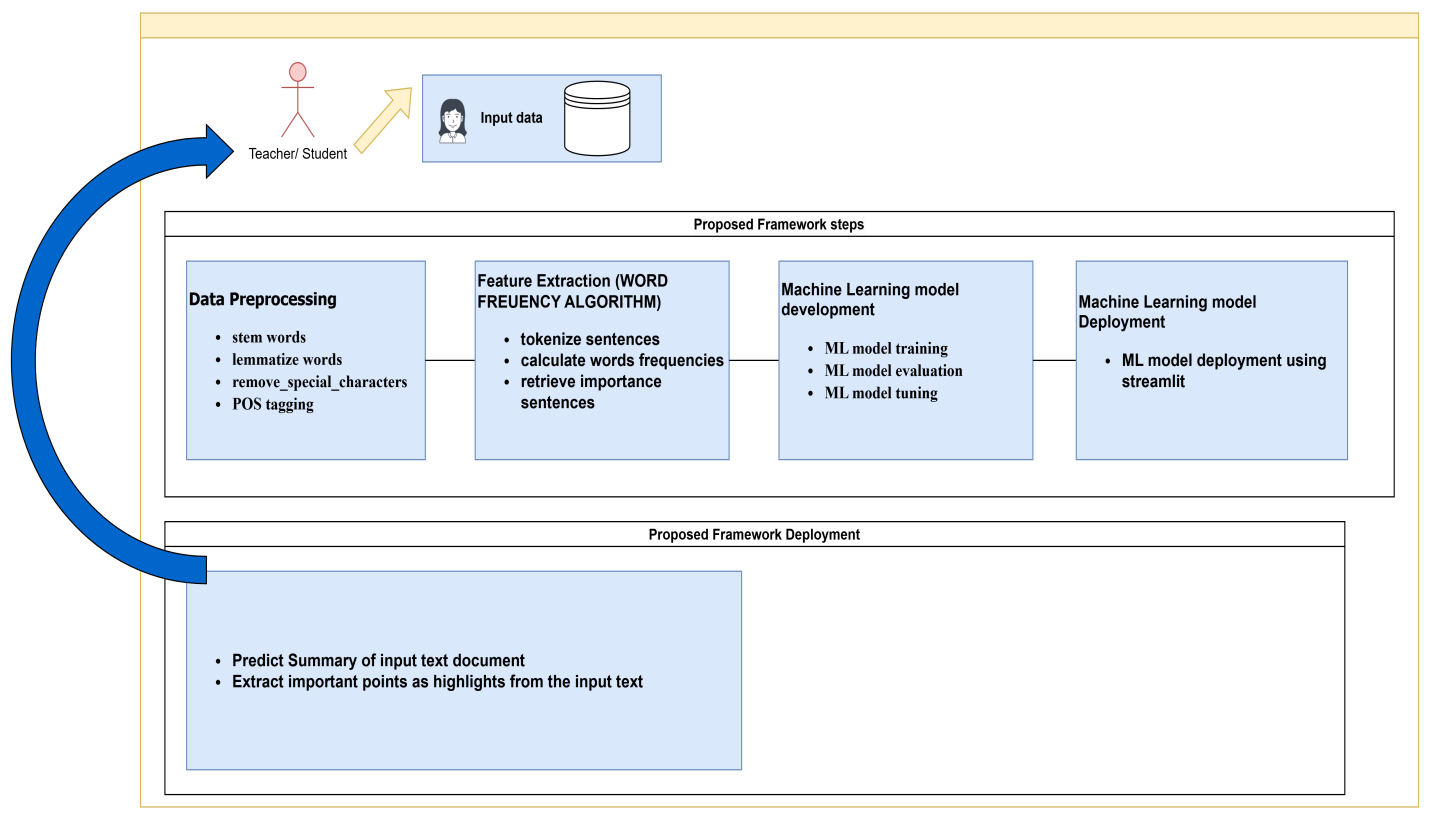
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ref. No.** | **Authors** | **Purpose** | **Model accuracy** | **Input Features** | **Year** | **Journal** |
| 1 | Zhenjie Zhao, YufangHou, DakuoWang, Mo YuFChengzhong Liu &Xiaojuan Ma | presented a novel framework in which questions from story books are generated through prior summarization | ROUGE L F1=41.78/38.29 | Educational Question-Answer Pairs, Annotated Questions, Control Signals, Text Summaries | 2022 | arXiv:2203.14187v1 [cs.CL] 27 Mar 2022 |
| 2 | Abdul GhafoorEtemad, Ali Imam Abidi, &Megha Chhabra | Comparison of two abstractive transformer models that are t5 and Bert by the use of the BBC News Summary dataset. | FT5 on Gigaword daraset  ROUGE1=43.02  ROUGE2=14.50  ROUGEL=37.43  ROUGELSUM=37.49 FT5 on Xsum dataset ROUGE1=30.91  ROUGE2=5.26  ROUGEL=20.85  ROUGELSUM=20.85 | Textual content of the articles from the Xsum dataset, which consist of articles from the BBC along with their corresponding summaries | 2021 | International Journal of Performability Engineering, vol. 17, no. 10, October 2021, pp. 900-906, DOI: 10.23940/ijpe.21.10. p8.900906 |
| 3 | Varun Deokar&Kanishk Shah | Comparison of Bart and T5 model for text-summarization of news articles was performed and a method for the automatic scraping of articles from google news using just keywords related to the article was proposed. | BART F1 score of about  33% and T5 model average F1 score of  about 26%. | requests, bs4[5], and newspaper packages to  automatically scrape the web and store the most relevant  article without any of the ads or irrelevant links. | 2021 | International Research Journal of Engineering and Technology (IRJET), Volume: 08, e-ISSN: 2395-0056 , p-ISSN: 2395-0072, www.irjet.net |
| 4 | James Mugi Karanja& Abraham Matheka | The authors aim to address the increasing challenge of information overload by developing a hybrid model that combines extractive and abstractive summarization techniques. | F1 SCORE ROUGE1=0.232166667,ROUGE2=0.04853333, ROUGEL=0.1416 | Text Content, Domain Information, Query, the structure of the document, including headers, figures, tables, and metadata, to aid in the analysis and summarization process, Language: | 2022 | Open Journal for Information Technology, 5(2), 65-80. ISSN (Online) 2620-0627 ▪ <https://doi.org/10.32591/coas.ojit.0502.03065k> |
| 5 | Mandy Guoy, Joshua Ainslie, David Uthus, Santiago OntañónJianmo Ni, Yun-Hsuan Sung &Yinfei Yang | to introduce a new Transformer architecture that allows for scaling both input length and model size simultaneously. The main goal is to explore the effects of this scaling on various natural language processing tasks, such as summarization and question answering. | LongT5 (xl - 16k input) =48.35 21.92 44.27 | **Scalability**: LongT5 allows for scaling both input length and model size simultaneously, **Transient Global Attention Mechanism**: LongT5 introduces a new attention mechanism called Transient Global (TGlobal) attention, which mimics the local/global attention mechanism of long-input transformers without requiring additional side-inputs, **Pre-training Strategy**: LongT5 adopts the pre-training strategy from PEGASUS, | 2022 | International Conference on Learning Representations (ICLR) |
| 6 | Shobhan Kumar, Arun Chauhan &Arun Chauhan | to summarize the input text by fine-tuned t5 model and also questions from this summarize output were generated | T5QG =40.96, 17.54, 19.21, 42.36  Contrastive\_T5QG= 42.04, 19.11, 20.07, 48.50 | Transformer Model, Pointer Net (Ptr-Net) Multi-Head Self-Attention Mechanism, Pointer Generator Network (PGN), Coverage Mechanism | 2022 | Proceedings of the 19th International Conference on Natural Language Processing (ICON) |
| 7 | Gagandeep Kaur &Amit Sharma | to introduce a novel framework called the SA model employing deep learning for CRS (SADL-CRS). This framework aims to enhance the efficiency of customer review summarization by leveraging sentiment analysis and review summarization techniques. | ROUGE L OF F1 CRS 19.31, 7.67, 18.71 | Review-related features (RRF), Aspect-related features (ARF), Hybrid feature vector (HF) | 2023 | International Journal of Electrical and Computer Engineering (IJECE) Vol. 14, No. 2, April 2024 |
| 8 | TohidaRehman, Suchandan Das, Debarshi Kumar Sanyal&SamiranChattopadhyay | The goal of the paper is to evaluate the performance of three different pre-trained models for abstractive text summarization. These models are:   1. Google/pegasus-cnn-dailymail 2. T5-base | ROUGE-1 F1: 36.45, ROUGE-2 F1: 15.92, ROUGE-L F1: 33.92, BLEU: 55.71, | Datasets: CNN-DailyMail, SAMSum, and BillSum, each with 2000 examples for testing.Pre-trained Models: google/pegasus-cnn-dailymail, T5-base, and facebook/bart-large-cnn.Token Limitations: | 2023 | arXiv:2303.12796v1 [cs.CL] 25 Feb 2023 |
| 9 | ZakariaeAlamiMerrouni , BouchraFrikh&BrahimOuhbi | The primary goal of the paper is to enhance the effectiveness of automatic text summarization (ATS) systems. This is achieved by introducing novel techniques for both extractive and abstractive summarization, focusing on improving the relevance and coherence of the generated summaries. The paper aims to tackle issues such as semantic redundancy and readability challenges in existing summarization methods | EXABSUMExtractive system achieved ROUGE-1 and ROUGE-2 scores of 0.601 and 0.451, respectively | Weighting Parameter (α): This parameter is crucial for determining the relevance of a term within a sentence, balancing statistical and semantic features. The optimal value found is α = 0.6​​.  Scoring Techniques: Various scoring techniques are combined to assess sentence relevance, including:  Term Relevance and Inverse Sentence Frequency (TR-ISF)  Sentence Position  Sentence Length  Sentence Resemblance to Title​ | 2023 | Journal of Big Data |
| 10 | AdhikaPramitaWidyassaria,b, SupriadiRustad a, GuruhFajarShidika,EdiNoersasongko Abdul Syukur a, AffandyAffandy a & De Rosal Ignatius Moses Setiadi | to present a novel approach to Automatic Text Summarization (ATS) that can generate both extractive and abstractive summaries. The paper proposes two distinct techniques: EXABSUMExtractive, which integrates statistical and semantic scoring methods to select and extract relevant, non-repetitive sentences from a text, and EXABSUMAbstractive, which employs a word graph approach (including compression and fusion stages) and re-ranking based on keyphrases to generate abstractive summaries using the source document as input . | performance on the DUC 2001 dataset with ROUGE-1 and ROUGE-2 scores of 0.480 and 0.208, respectively. For the DUC 2002 dataset, EXABSUMExtractive achieved ROUGE-1 and ROUGE-2 scores of 0.493 and 0.257, | Weighting parameter (α): This parameter governs the weight assigned to both the statistical feature (CHIR) and the semantic feature (SIM) within the hybrid weighting model. Different values of α (0, 0.2, 0.5, 0.6, 0.8, 1) were tested, with α = 0.6 yielding the most favorable outcomes .  TR-ISF measure: Utilized for sentence relevance detection, combining statistical and semantic features .  Sentence scoring techniques: Various methods to evaluate sentence relevance, including the use of statistical and semantic features . | 2022 | Journal of King Saud University – Computer and Information Sciences 34 (2022) 1029–1046 |
| 11 | Hans Christian, MikhaelPramodanaAgus&DerwinSuhartono | an extractive text summarization with TF-IDF method is used to build the summary | F-measure=0.666 | sentence based on its importance and given the value between zero and one | 2016 | Article in ComTech Computer Mathematics and Engineering Applications.DOI: 10.21512/comtech.v7i4.3746 |
| 12 | Marina Litvak, Natalia Vanetik, Sophie Krimberg | to propose a TF  -IDF weighing  method that helps to determine the most successful  candidate for the extractive summary among the  possible continuous document parts of the required  length. | F-measure=0.433 | text splitting, tokenization, special symbols  removal, removing of phone numbers, emails  etc. | 2020 |  |
| 13 | Winston C Zeng | to fine-tune Transformer-based natural language generation algorithms for USDA grains reports to benefit farmers, producers, and small businesses in the agricultural industry. The aim is to enhance the summarization of documents containing a significant amount of numerical information while maintaining grammatical structure and word flow. The research evaluates the performance of various NLP-trained models in summarizing documents with a focus on numerical data, with the ultimate objective of improving automated summarization processes for stakeholders in the agricultural sector. | Flan T5.Numerical=56%compression=62%....Bert .Numerical=17.25%compression=22% | textual data, pre-trained embeddings, attention mechanisms, and encoder-decoder architectures inherent in Transformer models to understand and generate summaries based on the input data. | 2024 | http://hdl.handle.net/10150/668771 |

Despite the growing interest in Text Summarization using an abstractive approach, there is a notable underrepresentation of studies focusing on hybrid approaches to text summarization that include both extractive and abstractive methods for summarizing text in an efficient way. The literature reveals a lack of consensus regarding the fine-tuning of T5 on a dataset that has already passed through some extractive technique, such as the word frequency algorithm or Term Frequency-Inverse Document Frequency (TF-IDF). Moreover, the scope of existing research is limited in its coverage of datasets from students' text books, as it may have the future implication of generating questions and making MCQs out of them. Furthermore, "A Hybrid Model for Text Summarization Using Natural Language Processing" by James Mugi Karanja & Abraham Matheka and "Fine-Tuned T5 for Abstractive Summarization" by Abdul Ghafoor Etemad, Ali Imam Abidi, and Megha Chhabra as highlighted in the literature review, are used as a model papers. The research gap that is founded in these papers is very low accuracy, and the papers lack a detailed comparison with existing state-of-the-art text summarization models. Addressing these accuracy-related limitations through improved model training, data preprocessing, and evaluation strategies could enhance the overall accuracy of the hybrid text summarization model proposed in these papers.

Despite significant advancements in text summarization techniques, existing methods face challenges in balancing accuracy, coherence, and contextual relevance. Traditional extractive methods like the Word Frequency Algorithm are efficient but often lack deep contextual understanding, while transformer-based abstractive models such as T5 and BART generate more coherent summaries but require fine-tuning for domain-specific accuracy. This research aims to address these issues by developing a hybrid summarization method that integrates the Word Frequency Algorithm with T5 and BART, enhancing the accuracy of these transformer models through fine-tuning on a self-created dataset, and deploying an application capable of generating summaries and highlighting key information from input texts.

**Methodology/Research Design**

This section outlines the methodology employed to address the research objectives outlined in the Research Objectives portion. The chosen methodology aims to provide a comprehensive understanding of “Automatic Text Summarization of Educational Books by Fine-Tuning of Abstractive Models.”.



ook of Grade 9th with Fine-Tuning Transformer-Based Natural Language Generation Algorithms” through a systematic approach.

Figure 1. Proposed framework

The first step in this research will involve the collection of an authentic and suitable dataset that fits for summarization purposes. Datasets can be chosen from newspapers, websites, articles, journals, etc. But the most suitable data to be summarized is the text from the student’s book. So, the dataset will contain the topics from educational books. A quantitative approach will be adopted to ensure a holistic investigation of the “Hybrid Model for Text Summarization Using an Extractive Abstractive Approach.”. Dataset will be passed from two techniques sequentially. Firstly, it will be passed from the extractive text summarization approach, which is a Word frequency algorithm. This technique is used to retrieve important sentences from the input text. Secondly, these retrieved sentences from the Word Frequency Algorithm will be passed from an abstractive approach, which is the T5 model. The T5 and BART models will be fine-tuned to achieve efficiency in performance and to increase their accuracy.

The research design encompasses an experimental research design, wherein quantitative data will be collected and analyzed. Since, fine-tuning a machine learning seq-to-seq model for text summarization will be performed, the study will fall under experimental research. As, the manipulation of an independent variable (the fine-tuning process) is observed, its effects on the dependent variable (the performance of the summarization model) are also involved.

Moreover, since working with a small dataset that has already undergone preprocessing using a Word Frequency Algorithm, our research design might be considered as pre-experimental or quasi-experimental. Pre-experimental designs lack a control group and may not have random assignment, which could be the case if our dataset has been pre-processed before fine-tuning.

**Conclusion:**

In conclusion, the hybrid approach for text summarization provides a systematic and rigorous approach to address the research objectives of this study. By employing a hybrid approach for text summarization through fine-tuning transformers, this study aims to generate a more precise and relevant summary. The deployment of the app will not only aid students, but also educators, in obtaining crucial information from lengthy text.

**References/Bibliography**

1. Zaware, S., Patadiya, D., Gaikwad, A., Gulhane, S., & Thakare, A., (2021). Text Summarization using TF-IDF and Textrank algorithm. *PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON TRENDS IN ELECTRONICS AND INFORMATICS (ICOEI).*
2. Elbarougy, R., Behery, G., & El Khatib, A. (2020). Extractive Arabic text summarization using modified PageRank algorithm. *EGYPTIAN INFORMATICS JOURNAL*, 21(2), 73-81.

1. Zhao, Z., Hou, Y., Wang, D., Liu, M. Y. & Ma, X. (2022, MARCH 27). Educational question generation of children storybooks via question type distribution learning and event-centric summarization. arXiv:2203.14187v1 [cs.CL]
2. Gopikakrishna, N. G., Sreenivasan, P., Chandran, V., K P, Y. K., Dev, S. S., & V V#, K., (2021). Comparative study on text summarization using nlp and rnn methods. International *JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT),* 9(13), [www.ijert.org](http://www.ijert.org)
3. Patil, H., Gaikwad, A., Dalave, M., Surywanshi, S., & Patil, P (2023). An automatic mcq & summary generation by using nlp. *INTERNATIONAL RESEARCH JOURNAL OF MODERNIZATION IN ENGINEERING TECHNOLOGY AND SCIENCE, 5,* **[www.irjmets.com](http://www.irjmets.com)**
4. El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert systems with applications, 165, 113679.
5. Etemad, A. G., Abidi, A. I. & Chhabra, M. (2021, OCTOBER). Fine-tuned t5 for abstractive summarization*. INTERNATIONAL JOURNAL OF PERFORMABILITY ENGINEERING*, 17(10),900-906. DOI: 10.23940/ijpe.21.10.p8.900906
6. Deokar, V. & Shah, K. (2021, SEPTEMBER 6). Automated text summarization of news articles*. INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET),* 8,1023–1033. https://doi.org/10.26615/978-954-452-092-2\_110
7. Karanja, J. M. & Matheka,A. (2022, NOVEMBER 28). A hybrid model for text summarization using natural language processing*. OPEN JOURNAL FOR INFORMATION TECHNOLOGY*, 5(2), 65-80. <https://doi.org/10.32591/coas.ojit.0502.03065k>
8. Guoy, M. , Joshua, A. , Uthus, D. , Ontañón, S. , Ni, J. , Sung, Y. & Yang, Y. (2022, MAY 3). Longt5: efficient text-to-text transformer for long sequences. https://github.com/google/flaxformer/tree/main/flaxformer/architectures/longt52https://github.com/google-research/longt5
9. Mengi, R. , Ghorpade, H., & Kakade, A., (2023, DECEMBER 5). Fine-tuning t5 and Robert a models for enhanced text summarization and sentiment analysis*. THE GREAT LAKES BOTANIST*. <https://www.researchgate.net/publication/376232167>
10. Kumar, S., Chauhan, A., & Kakade, P., (2022, DECEMBER). Augmenting ebooks with recommended questions using contrastive fine-tuned t5*. PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING (ICON),* 109 - 115, https://github.com/uci-soe/FairytaleQAData 2 https://ncert.nic.in/textbook.php
11. Kaur, G., & Sharma, A., (2024, APRIL). Automatic customer review summarization using deep learning-based hybrid sentiment analysis*. INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING (IJECE)* ,14(2), 2110-2125. DOI: 10.11591/ijece.v14i2.pp2110-2125
12. Liu, Y., Dou, Z., & Liu, P., (2021, APRIL 15). RefSum: Refactoring Neural Summarization. *arXiv:2104.07210v1 [cs.CL]*, https://github.com/yixinL7/ Refactoring-Summarization.http://explainaboard.nlpedia. ai/leaderboard/task-summ/index. php.
13. Rehman, T., Das, S., Sanyal, D. K., & Chattopadhyay, S., (2023 FEBRUARY 25). An analysis of abstractive text summarization using pre-trained models. arXiv:2303.12796v1 [cs.CL]
14. Merrouni, Z. A., Frikh, B., & Ouhbi, B., (2023)*. EXABSUM: a new text summarization approach for generating extractive and abstractive summaries. JOURNAL OF BIG DATA*, <https://doi.org/10.1186/s40537-023-00836->
15. Widyassaria, A. P., Rustad, S., Shidika, G. F., Noersasongko, E., Syukur, A., Affandy, A., & Setiadi, D. R. I. M., (2022). Review of automatic text summarization techniques & methods. *JOURNAL OF KING SAUD UNIVERSITY – COMPUTER AND INFORMATION SCIENCES* 34, 1029–1046
16. Abidin, A. Z. Z., Murdianingsih, Y., Suryadi U. T., & Setiyadi D., (2020). Text summarizing system of english subjects and text mining subjects*. JOURNAL OF CRITICAL REVIEWS,*Vol 7(5)
17. Christian, H., Agus, M. P., & Suhartono, D., (2016, DECEMBER). *Single document automatic text summarization using term frequency-inverse document frequency (tf-idf).COMTECH COMPUTER MATHEMATICS AND ENGINEERING APPLICATIONS*.DOI: 10.21512/comtech.v7i4.3746
18. Litvak, M., Vanetik, N., & Krimberg, N., (2020)*.*Summarization of financial documents with TF-IDF weighting of multi-word terms
19. Zeng, W. C., (2024, JANUARY 31). Fine-tuning Transformer-based Natural Language Generation Algorithms for USDA Grains Reports for Farmers, Producers, and Small Businesses. *THE UNIVERSITY OF ARIZONA USA* ,http://hdl.handle.net/10150/668771

**Gantt Chart (to be used as guideline)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Specific Objectives** | **Activities** | **Semester-I** | | | | **Semester-II** | | | |
| **Month-I** | **Month-II** | **Month-III** | **Month-IV** | **Month-I** | **Month-II** | **Month-III** | **Month-IV** |
|  | Literature’s Survey | Literature review |  |  |  |  |  |  |  |
| I |  | Literature review |  |  |  |  |  |  |  |
|  | Literature review | Decision about the  domain of research |  |  |  |  |  |  |
|  | Literature review | Decision about the  domain of research |  |  |  |  |  |  |
|  | Literature review | Decision about the  domain of research |  |  |  |  |  |  |
| II | Selection of Domain of research |  | Text Summarization | Literature Review about Text summarization using Extractive approach | Implementation of Word Frequency Algorithm in python |  |  |  |  |
|  |  |  | \_ | \_ | \_ |  |  |  |
|  |  |  |  |  | Literature Review about Text summarization using Abstractive approach | \_ | \_ |  |
| III | Implementation of Extractive Abstractive approach for text summarization |  |  |  |  | Synopsis writing | Synopsis writing | Synopsis submission |  |
|  |  |  |  |  | Implementation of WFA on selfcreated dataset | Implementation of WFA with T5 and Bert for finetuning | Fine tuning done with accuracy achieved |  |
|  |  |  |  |  | Implement coding on streamlit for making app | Start working on thesis | Start working on thesis |  |
| IV | Thesis writing & Submission |  |  |  |  |  | Thesis writing | Thesis writing |  |
|  |  |  |  |  |  |  | Thesis writing | Thesis submission | Thesis submission |

**Details of Funds/Expenditure (if applicable) *[Maximum 1 page]***

Provide detail of funds and expected expenditure on this research such as Consumable/Chemicals/field Survey/Transport etc.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. No.** | **Details** | **No(s)** | **Unit Rate** | **Total** |
| 1 | Consumables & Chemicals | \_\_ | \_\_ | \_\_ |
| 2 | Surveys or Transport | \_\_ | \_\_ | \_\_ |
| 3 | Contingencies | \_\_ | \_\_ | \_\_ |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| **Grand Total** | | | |  |

** Student Signature**

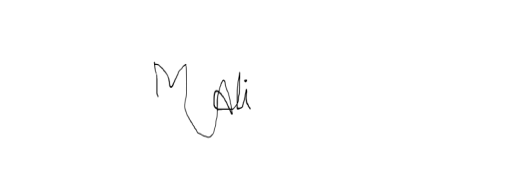
**Date:** 12/07/2024

**DECLARATION**

We hereby agree to supervise the research work as per above proposal/synopsis.



**Signature of Supervisor**

** **

**Signature of SC Member 1 Signature of SC Member 2**

**FOR VU THESIS SUPERVISOR USE ONLY**

**Date: 12/07/2024 Date: 12/07/2024**

*Note: Hard and soft copy of synopsis/research proposal must be submitted to secretary ASRB for final approval.*

**Profile of Supervisor**

**Name of Supervisor:** Mushtaq Hussain

**Designation:** Assistant professor

* Total No. of Impact Factor Research Publications during last 5 years: 5
* Total No. of Publications without Impact Factor during last 5 years: 7

|  |  |
| --- | --- |
| **Ongoing**  **Research students** | |
| Number of MS/M.Phil. students | Number of PhD students |
| 6 |  |

****

**Signature of Supervisor**

|  |  |
| --- | --- |
| Endst. No. \_\_\_\_\_\_\_\_\_\_\_ | Dated:12/07/2024 |

The Proposal entitled “Automatic Text Summarization of Educational Books by Fine-Tuning of Abstractive Models” duly recommended by the Graduate Research Committee (GRC) in its meeting held on \_\_\_\_\_\_\_\_\_\_ is forwarded to ASRB through the Dean of the Faculty for approval and allocation of funds (if requested).

|  |  |
| --- | --- |
|  | **Signature / Seal**  **Chairperson of the Department**  **Date: \_\_\_\_\_\_\_\_\_\_\_** |
| **Signature / Seal**  **Dean of the Faculty**  **Date: \_\_\_\_\_\_\_\_\_\_\_** | **Signature / Seal**  **Secretary ASRB**  **Date: \_\_\_\_\_\_\_\_\_\_\_** |