Domain-Adaptive Bangla Text Summarization: Harnessing Large Language Models in Closed-Set Contexts

Anusree Roy 2011364042
Department of
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
anusree.roy@northsouth.edu

Fahrin Hossain Sunaira 2031856042
Department of
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
fahrin.sunaira@northsouth.edu

Tateyama Orpa 2014239642
Department of
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
tateyama.orpa@northsouth.edu

Farasha Shamma Yussouf 2031250042
Department of
Electrical and Computer Engineering
North South University
Dhaka, Bangladesh
farasha.yussouf@northsouth.edu

***Abstract*—In the rapidly evolving digital landscape, the volume of textual information in the Bengali language has grown exponentially. Efficiently summarizing this vast amount of information is crucial for enabling users to quickly grasp key insights without sifting through extensive texts. With the help of large language models (LLMs) such as flanT5, mT5 this project aims to develop a domain-adapted Bengali text summarization model by training and finetuning on general and domain-specific datasets such as, state, international, and sports that enhance the model's ability to create more accurate summaries for specific content addressing the lack of effective text summarization models in a low-resource language like Bengali. The LLM models were trained on the XLSUM Bengali dataset for text summarization and then trained on source domains and fine-tuned on the target domains for domain adaptation. Here, the mT5 model outperformed the flanT5 model across most evaluation metrics, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.21, 0.09, and 0.20, respectively, alongside a BLEU score of 0.17 and BERTScore of 0.72. Furthermore, this project includes creating a custom dataset with manually annotated categories for domain adaptation purposes contributing to domain-specific Bengali data. Evaluation for domain adaptation using ROUGE, BLEU and BERTSCORE metrics demonstrated that flant5 XLSUM model outperformed other LLMs, achieving a ROUGE score of 0.79, BLEU score of 0.13, and BERTSCORE of 0.86, indicating its strong performance in generating high-quality Bengali text summaries for the state domain.**

***Index Terms*—Domain Adaptation, LLMs, ROGUE, BLEU, BERT, fine-tuning, zero-shot, closed-set**

I. Introduction

A. Background and Motivation

Text summarization is the procedure of shortening a large body of text into a condensed form while keeping the essential information within the overall text [1]. This approach is crucial in various aspects, including news reporting, where it helps readers quickly grasp the key points of the lengthy articles. Given that Bengali is spoken by over 250 million people worldwide [2], an efficient text summarization tool is essential, especially for news consumption. Summarizing Bengali news articles can significantly improve the accessibility and distribution of information, enabling readers to stay informed without needing to look through lengthy content [3]. Moreover, text summarization is a vital task in Natural Language Preprocessing (NLP), where the objective is to condense long texts into shorter and informative summaries. However, there is a noticeable lack of comprehensive datasets for the Bengali language. Several remarkable research publications [4] [5] on Bengali text summarization utilize the same few public datasets that are available. The absence of such datasets hampers the development and refinement of summarization techniques tailored for Bengali text summarization [6].

Recent studies show the significant advancements made by Large Language Models (LLMs) in various natural language processing tasks, including abstractive text summarization. Research has also examined the domain adaptation potential of LLMs for summarization tasks, analyzing how these models perform when adapted to specific domains [7].

Domain Adaptation is a special case of transfer learning where the task remains unchanged, but the domains differ. It aims to improve model performance on a target domain with insufficient annotated data by using knowledge from a related source domain with adequate labeled data. In contrast to transfer learning, which also involves task and/or domain changes, domain adaptation focuses on aligning distributions between the source and target domains to enhance performance. It allows models to adapt to new domains where annotated data is limited by leveraging knowledge from related domains [8]. Also, domain adaptation is a critical approach for enhancing summarization quality, it involves fine-tuning a model on a specific domain's data, which allows the model to better understand domain-specific terminology and context. This process allows models to better capture the vocabulary, structure, and context that are unique to particular domains, such as news, healthcare, or finance [9]. The significant progress of Large Language Models (LLMs) in a variety of natural language processing tasks has been well-documented, including abstractive text summarization, which involves generating a condensed version of the most relevant information from a document [10]. Research on domain adaptation in LLMs for summarization has been growing, but most studies have focused on single domains, such as news articles or clinical reports. There is a need for further research across multiple domains to better understand how these models adapt to different target areas [8].

When LLMs are applied to specific domains, they require corresponding domain-specific knowledge bases for training. This leads to a limitation where a model trained in one domain struggles to generalize to others, causing resource inefficiency. The issue arises from the disparity between the training and target domain data distributions. This adaptation is especially crucial when working with specialized topics such as sports, international, or states, where general-purpose models may struggle to generate relevant summaries. Therefore, methods to bridge this gap and enhance the model's adaptability and efficiency in summarization tasks are essential. Domain adaptation seeks to train a model using multiple source domains, allowing it to generalize effectively to new, unseen domains. As a result, improving domain adaptation performance is a critical goal for large-scale models to enhance their effectiveness in downstream tasks. Exploring the factors that influence domain adaptation performance is an important area of focus [9].

In response to the challenges in low-resource domain scenarios, particularly for abstractive summarization, domain adaptation methods have emerged as vital solutions. These approaches enable models to quickly adapt to target domain tasks despite the scarcity of labeled data. Yet, very few studies have applied domain adaptation techniques to low-resource scenarios for the abstractive summarization task. Such methods reduce the number of training samples in various domains to create low-resource scenarios, testing the adaptability and performance of summarization models under constrained conditions [11] .

The concept of domain adaptation and fine-tuning-based training for abstractive text summarization can be compared under different scenarios such as monolingual, cross-lingual, and multilingual summarization tasks. In domain adaptation, fine-tuning achieves better results when enough training data is available and it's also more effective when there are significant differences between the domains [12]. Accordingly, this approach is employed in this project.

There can be some research gaps [13] in the domain-specific abstractive text summarization such as handling large input texts can be challenging and therefore using an efficient transformer is necessary. To ensure good accuracy, it is suggested to apply advanced evaluation metrics such as BERTscore, etc. so that we can easily detect and fix model hallucinations. Lastly, the problems with models being trained on general-purpose data have also been addressed as they perform poorly with specialized content because they lack domain-specific knowledge. Thus, fine-tuning pre-trained models on domain-specific data can be a potential solution to this issue.

B. Purpose and Goal of the Project

The objective of this project is to develop a domain-adapted Bengali text summarization model leveraging large language models (LLMs) such as mT5 and flanT5. The project addresses the lack of effective text summarization models in a low-resource language like Bengali which adds uniqueness. A custom dataset comprising 21,512 rows was created by scraping various Bengali news articles with each data manually categorized into seven distinct domains such as sports, states, entertainment, international, education, economy, and technology. Among these, the two LLMs were fine-tuned across 3 domains with the highest rows: states, international, and sports. Thus, all of these points mentioned above collectively contribute to the novelty of this project.

C. Organization of the Report

The rest of the paper is divided into seven chapters. Chapter II provides the literature review where related studies on domain adaptive Bengali text summarizations are discussed. The following chapter III addresses the methodology. Throughout the research process, numerous experiments were carried out, and the results have been thoroughly discussed in chapter IV, while in chapter V the impacts of the project on various sectors have been outlined. Chapter VI focuses on the project planning and budget, illustrated using a Gantt chart and a budget table. Finally, chapter VII concludes the paper by highlighting the project's limitations and providing insights for future work.

II. Literature Review

A. Existing Research and Limitations

Text Summarization using Large Language Models: This paper [14] embarks on an exploration of text summarization with a diverse set of LLMs, including MPT-7b- instruct, falcon-7b-instruct, and OpenAI Chat-GPT textdavinci-003 models. The primary objective is to provide a comprehensive understanding of the performance of Large Language Models (LLMs) like MPT-7b-instruct, falcon-7b-instruct, and OpenAI ChatGPT textdavinci-003 when applied to different datasets like XSUM and CNN-DailyMail News Text Summarization datasets. From their experiment, text-davinci-003 outperformed the others on both datasets. Borah et al. [15] obtained the T5 model's effectiveness in abstractive text summarization on datasets like CNNDM, MSMO, and XSUM. Their ROUGE and BLEU scores demonstrated the model's superior capability in generating concise summaries, particularly excelling with the MSMO dataset. The paper by Lewis et al. [16] introduces the BART model which outperforms existing models in text summarization tasks, achieving top ROUGE scores of 44.16 on the dataset CNN/DailyMail and 45.14 on XSum. Moreover, the recent study by Song et al. [17] makes use of FineSurE, which is a novel evaluation framework that employs large language models (LLMs) to assess summarization quality across linguistic, content, and factual dimensions. The system outperforms current approaches in both general and domain-specific summarization tasks, achieving state-of-the-art accuracy with correlations as high as 0.92 with human assessments on benchmark datasets. The research [18] focuses on uncertainty propagation in text summarization using large language models (LLMs) like GPT-4 and T5. Here, they implement model-based uncertainty metrics in their framework to quantify and align uncertainty transfer. Demonstrated on benchmark datasets like CNN/DailyMail and XSum, their method significantly reduces uncertainty mismatch, with alignment measures demonstrating improvements of up to 15% above baselines.

Text Summarization in Bengali: The study [4]introduces an approach where a single text is first summarized using four pre-trained Bengali text summarization models, namely - mT5 XLSum, mT5 CrossSum, Scibert uncased, and mT5 by Shahidul. Then its respective human-written reference summary is taken. The proposed approach for choosing the best summary involves evaluating the similarity of all candidate summaries with the reference summary (human-written). The top-ranked summary is picked based on the summary similarity assessment, summary similarity matrix calculation, graph building, and ranking. Their one limitation is that the mt5 XLSum model performs better than other models on the Bengali Text Summarization dataset but not as well as on the XLSUM dataset. The paper by Kowsher et al. [19] introduces a monolingual BERT model for the Bengali language called Bangla-BERT. They constructed a Bengali language model dataset, BanglaLM. This paper resolves the mBERT's limitation for Bengali trained on limited and more structured data only and mixed weights issues among 104 languages. The model outperformed other models and surpassed all prior state-of-the-art results by 3.52%, 2.2%, and 5.3%. In 2020, Prithwiraj Bhattacharjee et al. [5]developed a seq2seq LSTM model with a local attention mechanism for Bengali abstractive news summarization which effectively generated coherent and contextually rich summaries from a large dataset of over 19,000 articles. Next, Chowdhury et al. [20] developed a graph-based unsupervised approach, leveraging the seq2seq architecture to enhance the abstractive summarization of Bengali texts, which are notably low-resource, demonstrating significant potential for natural language processing applications in under-resourced languages. Jahan et. al [21]proposed a unique study in the field of bengali text summarization where they used a lexicon-based method. The method effectively summarizes text and extracts important sentences by using rule-based algorithms and a sentiment lexicon. When assessed on a dataset of Bengali text, the method's summarization accuracy was 78%, indicating that it is capable of generating meaningful and eloquent summaries without the need for machine learning models.

Bengali Text Summarization using LLMs: The recent study by Rony et al. [22] evaluates the performance of large language models (LLMs), including GPT-4 and mT5, for Bangla text summarization. Using Bengali-specific datasets, the study compares these models and demonstrates how GPT-4 outperformed the others, achieving the highest ROUGE-L score of 82%. The results illustrate how well LLMs can handle low-resource languages like Bangla, but they also stress how crucial dataset quality and fine-tuning are to getting the best results. Next, Lora et al. [23] explores ConVerSum, a framework utilizing contrastive learning for cross-lingual summarization with a focus on Bengali. To align semantically similar words across languages, the technique incorporates pre-trained models like mBART and XLM-R for representation learning in conjunction with a contrastive loss. ConVerSum outperforms current techniques in terms of accuracy and fluency, especially in situations where data is scarce, with a ROUGE-L score of 84% for Bengali cross-lingual summary tasks when analyzed on benchmark datasets. Furthermore, another study [24] introduces SumTra, a differentiable pipeline designed for few-shot cross-lingual summarization, with a focus on Bengali. Effective summarization in low-resource languages is made possible by the framework's combination of few-shot learning with pre-trained models like T5 and mBART. With a ROUGE-L score of 85% for Bengali text summarization, the method showed promise for scalable cross-lingual applications and performed well even with little training data. Another work in the field of Bengali text summarization using LLMs is conducted by Zehady et al. [25] where they use BongLLaMA, a variant of the LLaMA model fine-tuned specifically for the Bengali language. The model's abilities are demonstrated in a variety of NLP tasks, such as summarizing Bangla text. The model outperformed other LLMs like mBERT and conventional methods, achieving a ROUGE-L score of 88% on Bengali summarizing tasks using BongLLaMA.

Domain Adaptation: The study [26] focuses on unsupervised domain adaptation—a machine learning approach that addresses the challenge of adapting models trained on a source domain to a target domain with different distributions, without requiring labeled data in the target domain. It also describes different sets of domain adaptation like- closed-set, open-set, partial, and universal domain adaptation. Next, Busto et al. [27] explore the problem of domain adaptation in scenarios where the target domain contains categories not present in the source domain. This approach, known as Open Set Domain Adaptation (OSDA), is particularly valuable in real-world applications where data in the target domain includes instances of unknown classes that aren't in the source domain. The authors conducted extensive experiments on datasets like the Office dataset (images of office supplies), the Caltech dataset, and video datasets for action recognition. Their model, referred to as ATI- (Assign-and-Transform-Iteratively with Outlier handling), showed strong performance on open-set tasks by accurately identifying unknown categories in the target domain. Additionally, the research [28] addresses the limitations of conventional domain adaptation approaches when dealing with partial transfer learning scenarios. In these scenarios, only a subset of classes from the source domain overlaps with the target domain, making it essential to ignore irrelevant source classes to avoid negative transfer. The model is evaluated on datasets like Office-31, Caltech-Office, and ImageNet-Caltech, covering diverse transfer learning scenarios. They introduce the Selective Adversarial Network (SAN), a model that selectively aligns the source and target domains by focusing only on shared classes while disregarding irrelevant source classes. The paper [29] introduces a new approach to domain adaptation called Universal Domain Adaptation (UDA). This framework is designed for scenarios where the relationship between the label sets of the source and target domains is unknown, enabling the model to handle a mix of shared and private classes in both domains. They carried UDA on datasets like- Office-31, Office-Home, VisDA2017, ImageNet-Caltech.

Domain Adaptation in the field of Text Summarization: The research [7] explores the domain adaptation capabilities of various large language models (LLMs) in text summarization. It evaluates 11 models, including conventional encoder-decoder models and LLMs of various sizes, across three domains: scientific, medical, and governmental. The evaluation is done using two methods: fine-tuning and in-context learning (ICL). The authors present a domain adaptation evaluation suite called AdaptEval, which includes a domain-specific benchmark and metrics such as ROUGE, BERTScore, domain vocabulary overlap (DVO), token distribution shift, and G-eval. The results show that LLMs perform comparably well in the ICL setting, fine-tuning models achieve better automatic evaluation scores. In addition to this, the study [30] explores various strategies for fine-tuning large language models (LLMs) to enhance their domain adaptation capabilities. They examine how different fine-tuning techniques, such as CPT and SFT, impact LLM performance in specialized domains. Results reveal that models fine-tuned and merged using SLERP and preference optimization strategies achieve higher accuracy in domain-specific tasks. Moreover, Zhong et al. [31] introduces MTL-DAS, a model for automatic text summarization focused on domain adaptation, by using transformer-based models that have already been trained, such as BERT, and fine-tuning them for particular domains. In order to improve summarizing performance, the authors provide a multi-task learning (MTL) approach that uses domain-specific data to modify summarization models to particular domains. When tested on domain-specific datasets, MTL-DAS outperforms conventional summarization models and significantly improves ROUGE-L scores, with the highest result being approximately 90%. Other significant approaches in the field of domain adaptation for text summarization include the study by Mao et al. [32] where they introduce CiteSum which is a framework for scientific extreme summarization. This framework uses citation text to direct the summarizing process and may adapt to certain domains with minimal supervision. A small set of labeled data from scientific citation datasets, including publications from repositories like arXiv and PubMed, has been utilized by the authors to fine-tune transformer-based models like BERT and T5. CiteSum performs impressively with ROUGE-L score reaching up to 90% proving its capacity to produce concise and insightful summaries while maintaining accuracy in domain adaptation.

B. Comparative Analysis of Related Works

A comparative analysis of some of the related works on Bengali Text Summarization using LLMs and Domain Adaptation as per provided on the literature review section are illustrated in Table I.

III. Methodology

A. Problem Formulation

Fig. 1 shows the primary steps taken in this research. To formally describe the methodology, let X_i denote the input Bangla text sequence, where i refers to the sample index. After preprocessing (including normalization, tokenization, and padding), X_i is transformed into a tokenized sequence T_i of fixed length L. This sequence is embedded into a dense vector $E_i \in \mathbb{R}^d$, where d is the embedding dimension, with positional encodings P_i added to maintain sequence order. The combined representation $E_i + P_i$ is passed through the encoder, generating semantic latent representations $h_i \in \mathbb{R}^m$. These latent representations h_i are then processed by the decoder to predict the tokenized summary $\hat{Y}_i = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_L\}$, where $\hat{y}_k$ represents the k-th token in the predicted summary. For domain adaptation, the latent representation h_i is used to compute the domain classification loss L_D by comparing the predicted domain label $\hat{Y}_{\text{domain}}$ against the ground truth domain label Y_{domain}. The gradient reversal layer modifies h_i to produce domain-invariant embeddings $\tilde{h}_i \in \mathbb{R}^m$, which are used to calculate the domain confusion loss L_{DC}. These

TABLE I
A Comparative Analysis of Some of the Related Works

REF	Subject	Dataset used and size	Model used and evaluation metrics	Advantages	Limitations
[14]	Text summarization using different LLM models.	The CNN/Daily Mail 3.0.0 dataset and XSum dataset containing 201k training samples.	LLM models used: mpt-7b-instruct, falcon-7b-instruct, and text-davinci-003, evaluated by BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore.	Use of larger models and fine-tuning on specific domains.	Evaluation is limited to only two datasets, which may not fully represent the performance of the models.
[15]	Comparative analysis of T5 model for abstractive text summarization.	Used CNN/DM, MSMO, and XSum datasets.	T5 model for abstractive summarization, evaluated using ROUGE and BLEU scores.	T5 model is adaptable for multiple NLP tasks.	Model has high computational demands and potential inconsistencies in summaries.
[4]	Bengali text summarization using rank-based approach.	XL-Sum multilingual dataset (10126 rows). 'Bangla Text Summarization dataset' with 80.3k data.	Four models used: mt5 XL-Sum, mT5 CrossSum, scibert uncased, and mT5. Evaluated using BLEU, ROUGE, BERTScore, WIL, and METEOR.	Compares and shows how each of the models performs.	Outputs were not consistent for the two different datasets used.
[19]	Monolingual BERT model for the Bengali language called Bangla-BERT.	They constructed a Bengali language model dataset, BanglaLM.	Bangla-BERT model has been trained on Transformer-based BERT architecture and it's evaluated using accuracy, precision, recall, F1-score, AUC and Hamming loss.	Obtains better accuracy than mBERT and other traditional models in Bangla NLP tasks like sentiment analysis, NER,binary and multi-label classification.	High computational cost for training and inference. Requires large datasets for domain specific fine-tuning.
[22]	Evaluating large language models for summarizing Bangla texts.	Uses BANS and BNLPC datasets, and then uses 50 examples from each validation set for human evaluation.	Five models used: GPT-3.5, GPT-4, OPT, LLaMA2, PaLM-2. Evaluated using BLEURT, ROUGE, BERTScore, BARTScore, and METEOR.	Compares five different models, and GPT-4 shows the best results.	Uses smaller sample size than other text summarization datasets.
[7]	Evaluating large language models for summarizing texts from different domains.	For the science domain, 215.8k data were used from the PubMed dataset. For the government domain, 193k data were used from the GovReport dataset.	Models used: ChatGPT, GPT-4, Falcon, LLaMA2, Vicuna, Mistral, and PEGASUS. Evaluated using ROUGE, BERTScore, and DVO.	Performs summarization on three different domains.	For GPT-4, only 25 random samples due to high cost.

REF	Subject	Dataset used and size	Model used and evaluation metrics	Advantages	Limitations
[16]	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.	GLUE benchmark, SQuAD 1.1 and 2.0, CNN/Daily Mail, XSum, MNLI, and WMT16 are the datasets used here.	BART model was evaluated with ROUGE (R1, R2, RL), BLEU, Perplexity (PPL), Accuracy, F1 Score, MCC, and MET for tasks such as summarization, classification, and question answering. Comparison with other models such as ELMo, GPT, XLNet, UniLM, and MASS.	It combines bidirectional encoding and autoregressive decoding, which allows flexible masking.	High computational costs, large dataset dependency, and less effective for loosely aligned tasks like ELIs.
[5]	Bengali abstractive summarization model: An Attention-Based Approach.	19,096 articles were collected from a Bengali corpus (BANPS). Annotated data are split equally on both encoder and decoder.	Model used: LSTM-based architecture evaluated using ROUGE, BLEU scores, and user ratings.	Generates human-like summaries with scores that outperform heuristic methods in fluency.	Struggles with long words, large repetitions, and truncation due to the length-based training.
[24]	SUMTRA: A Differentiable Pipeline for Few-Shot Cross-Lingual Summarization	Datasets used: CrossSum and WikiLingua with high, medium, and low-resource languages and its sizes varies depending on the language pairs.	SUMTRA: combines a monolingual summarizer and a translator(mBART-50 variants) and the evaluation metrics used were mROUGE (ROUGE-1,ROUGE-2, ROUGE-L) and BERTSCORE.	Shows strong zero-shot and few-shot performance, modular and end-to-end differentiable, effective for low resource languages.	Depends on high-quality summarization and translation modules, needs higher memory and compute needs and it's limite to English-to-many setups.
[28]	Partial Transfer Learning with Selective Adversarial Networks	Office-31 consisting of 4,652 images,31 categories, Caltech-Office(10 categories) and ImageNet-Caltech (84 categories) datasets were used.	Selective Adversarial Networks (SAN) was used as the model and it was evaluated using classification accuracy on transfer tasks.	Filters irrelevant classes to avoid negative transfer and achieves state-of-the-art performance.	High computational cost and it depends on reliable target data probabilities.

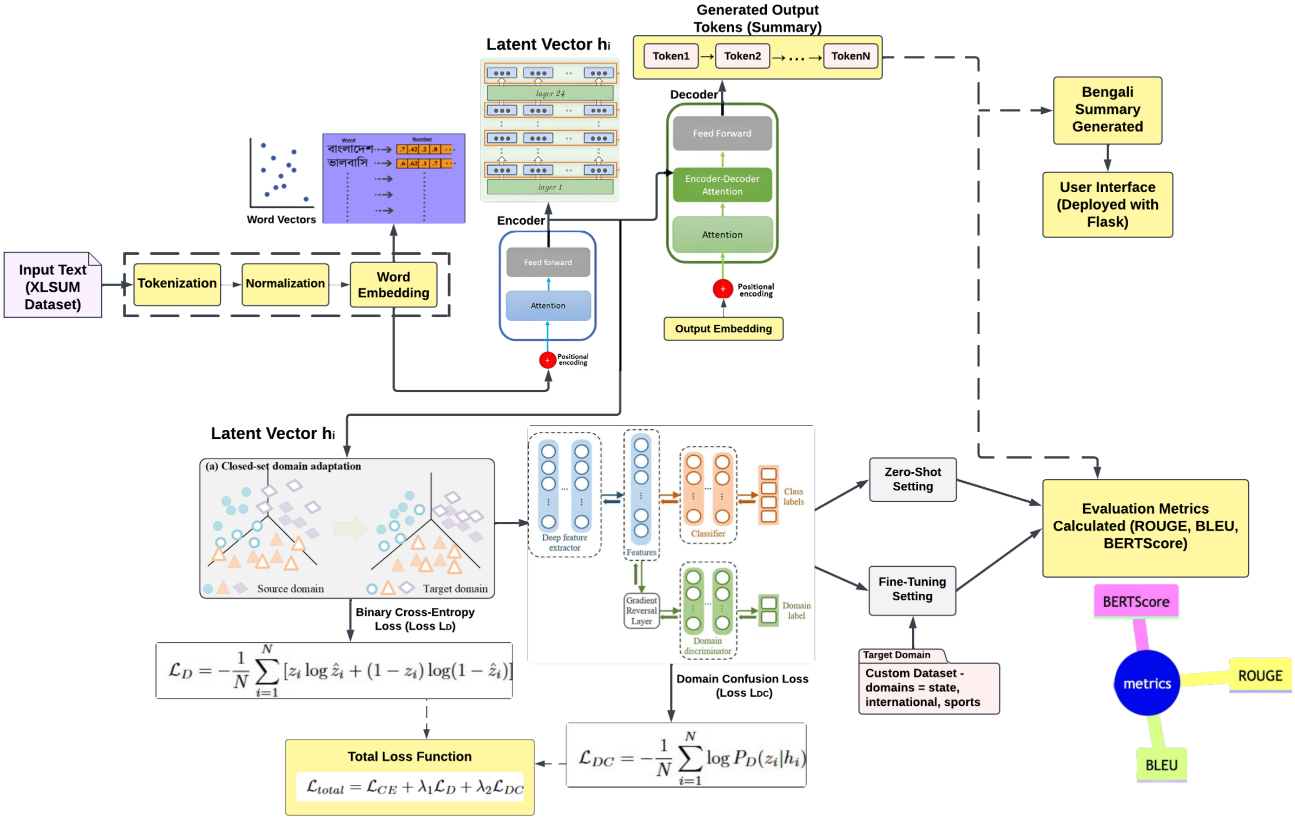

Generated Output
Tokens (Summary)
Token1
Token2
TokenN
Latent Vector hi
layer 24
layer 1
Decoder
Feed Forward
Encoder-Decoder Attention
Attention
Positional encoding
Output Embedding
Encoder
Feed forward
Attention
Positional encoding
Bengali Summary Generated
User Interface (Deployed with Flask)
Word Vectors
Word
Number
বাংলাদেশ
ভালবাসি
Input Text (XLSUM Dataset)
Tokenization
Normalization
Word Embedding
Latent Vector hi
(a) Closed-set domain adaptation
Source domain
Target domain
Deep feature extractor
Features
Classifier
Class labels
Gradient Reversal Layer
Domain discriminator
Domain label
Zero-Shot Setting
Fine-Tuning Setting
Target Domain
Custom Dataset - domains = state, international, sports
Evaluation Metrics Calculated (ROUGE, BLEU, BERTScore)
BERTScore
metrics
ROUGE
BLEU
Binary Cross-Entropy Loss (Loss LD)
$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N}[z_i \log \hat{z}_i + (1 - z_i)\log(1 - \hat{z}_i)]$
Domain Confusion Loss (Loss LDC)
$\mathcal{L}_{DC} = -\frac{1}{N}\sum_{i=1}^{N}\log P_D(z_i|h_i)$
Total Loss Function
$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_D + \lambda_2\mathcal{L}_{DC}$

Fig. 1. Workflow of the project approach

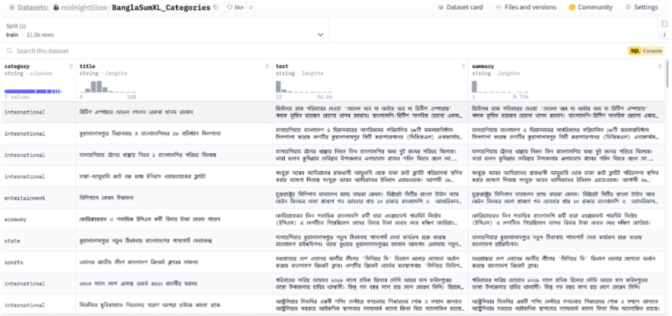
embeddings are optimized to minimize the domain-specific variance while retaining task-relevant information. The total loss function is given as:

$$L_{\text{total}} = L_{\text{CE}} + \lambda_D L_D + \lambda_{DC} L_{DC},$$

where L_{CE} measures the error between the predicted summary $\hat{Y}_i$ and the ground truth Y_i. The scaling factors λ_D and λ_{DC} balance the contributions of the domain adaptation losses. During evaluation, the test and validation inputs X_i are processed through the same pipeline. Performance metrics, such as BLEU, ROUGE, and BERTScore, are computed to evaluate the quality of the predicted summaries. The final output $\hat{Y}_i$ may undergo optional post-processing before deployment.

B. Custom Dataset Preparation

For domain adaptation, a dataset of 21,512 rows was created by web scraping news articles from online newspapers including “The Daily Star”, “Prothom Alo”, “Jugantor” and “Ittefak”. For the web scraping process Python’s Beautiful Soup library was leveraged. Beautiful Soup is a package in Python that parses HTML and XML documents. It is used to generate a parse tree for documents for extracting data from HTML needed for web scraping. In the Python code, the class names of the HTML pages for different tags - `<h1>` and `<p>` were assigned and according to that class name the data under it was collected, therefore, for different online newspapers, different class names that were assigned on that page had to be used. The title of the news article was labelled as ‘title’ column, the headline at the top of a newspaper containing the summarized information was labelled as the ‘summary’ column and the whole text part describing news was categorised as the ‘text’ column. A total of 22,100 rows were collected. After that, the dataset was divided into 7 categories - sports, international, state, entertainment, economy, education, and technology which were manually annotated. Following this, the duplicates and null values were removed from the dataset. so, after conducting Exploratory Data Analysis (EDA), the dataset was refined to 21,512 rows. Fig. 2 illustrates a sample of the custom dataset curated, which has 4 columns - category, title, text, and summary. For EDA, the ‘Pandas’ library of Python was utilized, which is a data manipulation tool. Additionally, after dropping the duplicate and null values, data sampling was done for the domain adaptation part, due to the state category having too many values compared to other categories for both the custom dataset (target domain) and source domain dataset. For resampling, the scikit-learn library called resample was used. Finally, the resampled data of both source and target domain were obtained with the following distribution: 42.9% for state, 35.7% for international, and



Datasets: midnightGlow/BanglaSumXL_Categories ♡ like 0 · Dataset card · Files and versions · Community · Settings

Split (1)
train · 21.5k rows

Search this dataset · SQL Console

category string · classes 7 values	title string · lengths 4 148	text string · lengths 22 34.6k	summary string · lengths 1 8.73k
international	ব্রিটিশ এম্পায়ার মেডেল পেলেন রেহানা খানম রহমান	ব্রিটেনের রাজ পরিবারের দেওয়া 'মেডেল অব দ্য অর্ডার অব দ্য ব্রিটিশ এম্পায়ার' পদকে ভূষিত হয়েছেন রেহানা খানম রহমান। বাংলাদেশি-ব্রিটিশ নাগরিক রেহানা একজ…	ব্রিটেনের রাজ পরিবারের দেওয়া 'মেডেল অব দ্য অর্ডার অব দ্য ব্রিটিশ এম্পায়ার' পদকে ভূষিত হয়েছেন রেহানা খানম রহমান। বাংলাদেশি-ব্রিটিশ নাগরিক রেহানা একজ…
international	কুয়ালালামপুরে মিয়ানমার ও বাংলাদেশিদের ১৮ প্রতিষ্ঠান সিলগালা	মালয়েশিয়ায় বাংলাদেশ ও মিয়ানমারের নাগরিকদের পরিচালিত ১৮টি ব্যবসাপ্রতিষ্ঠান সিলগালা করেছে দেশটির কুয়ালালামপুর সিটি করপোরেশনের (ডিবিকেএল) এনফোর্সম…	মালয়েশিয়ায় বাংলাদেশ ও মিয়ানমারের নাগরিকদের পরিচালিত ১৮টি ব্যবসাপ্রতিষ্ঠান সিলগালা করেছে দেশটির কুয়ালালামপুর সিটি করপোরেশনের (ডিবিকেএল) এনফোর্সম…
international	মালয়েশিয়ায় ট্রেনের ধাক্কায় নিহত ২ বাংলাদেশির পরিচয় মিলেছে	মালয়েশিয়ায় ট্রেনের ধাক্কায় নিহত তিন বাংলাদেশির মধ্যে দুই জনের পরিচয় মিলেছে। তারা হলেন কুমিল্লার দেবিদ্বার উপজেলার এলাহাবাদ গ্রামের শহিদ মিয়ার ছেলে মো…	মালয়েশিয়ায় ট্রেনের ধাক্কায় নিহত তিন বাংলাদেশির মধ্যে দুই জনের পরিচয় মিলেছে। তারা হলেন কুমিল্লার দেবিদ্বার উপজেলার এলাহাবাদ গ্রামের শহিদ মিয়ার ছেলে মো…
international	ঢাকা-আবুধাবি রুট বন্ধ হচ্ছে ইত্তিহাদ এয়ারওয়েজের ফ্লাইট	সংযুক্ত আরব আমিরাতের রাজধানী আবুধাবি থেকে ঢাকা রুটে ফ্লাইট পরিচালনা স্থগিত করার ঘোষণা দিয়েছে সংযুক্ত আরব আমিরাতের ইত্তিহাদ এয়ারওয়েজ। আগামী ১৯…	সংযুক্ত আরব আমিরাতের রাজধানী আবুধাবি থেকে ঢাকা রুটে ফ্লাইট পরিচালনা স্থগিত করার ঘোষণা দিয়েছে সংযুক্ত আরব আমিরাতের ইত্তিহাদ এয়ারওয়েজ। আগামী ১৯…
entertainment	মিশিগানে জেমস উন্মাদনা	যুক্তরাষ্ট্রের মিশিগানে মাতালেন ব্যান্ড তারকা জেমস। ডিট্রয়েট সিটির বাংলা টাউন নামে জেইন ফিল্ডের মেলা প্রাঙ্গণে গত রোববার প্রায় ২০ হাজার বাংলাদেশি ও আমেরিকান…	যুক্তরাষ্ট্রের মিশিগানে মাতালেন ব্যান্ড তারকা জেমস। ডিট্রয়েট সিটির বাংলা টাউন নামে জেইন ফিল্ডের মেলা প্রাঙ্গণে গত রোববার প্রায় ২০ হাজার বাংলাদেশি ও আমেরিকান…
economy	কোরিয়াফেরত ৩ শতাধিক ইপিএস কর্মী বিমার টাকা ফেরত পাবেন	কোরিয়াফেরত তিন শতাধিক বাংলাদেশি কর্মী যারা এমপ্লয়মেন্ট পারমিট সিস্টেম (ইপিএস) এ দেশটিতে গিয়েছিলেন তাদের বিমার টাকা ফেরত দেবে দক্ষিণ কোরিয়া।…	কোরিয়াফেরত তিন শতাধিক বাংলাদেশি কর্মী যারা এমপ্লয়মেন্ট পারমিট সিস্টেম (ইপিএস) এ দেশটিতে গিয়েছিলেন তাদের বিমার টাকা ফেরত দেবে দক্ষিণ কোরিয়া।
state	কুয়ালালামপুরে নতুন ঠিকানায় বাংলাদেশের পাসপোর্ট সেবাকেন্দ্র	মালয়েশিয়ার কুয়ালালামপুরে নতুন ঠিকানায় পাসপোর্ট সেবা কার্যক্রম শুরু করেছে বাংলাদেশ হাইকমিশন। আজ বুধবার কুয়ালালামপুরের জালান আমপাং এলাকায় নতুন…	মালয়েশিয়ার কুয়ালালামপুরে নতুন ঠিকানায় পাসপোর্ট সেবা কার্যক্রম শুরু করেছে বাংলাদেশ হাইকমিশন।
sports	ওমানের জাতীয় লীগে বাংলাদেশ ক্রিকেট ক্লাবের সাফল্য	মধ্যপ্রাচ্যের দেশ ওমানের জাতীয় লীগের 'ডিভিশন ডি' বিভাগে খেলার যোগ্যতা অর্জন করেছে বাংলাদেশ ক্রিকেট ক্লাব। দেশটির ক্রিকেট বোর্ডের ব্যবস্থাপনায় 'ডিভিশন ডিভিশ…	মধ্যপ্রাচ্যের দেশ ওমানের জাতীয় লীগের 'ডিভিশন ডি' বিভাগে খেলার যোগ্যতা অর্জন করেছে বাংলাদেশ ক্রিকেট ক্লাব।
international	২০২৩ সালে দেশে এসেছে রেকর্ড ৪০০২ প্রবাসীর মরদেহ	পরিবারের দারিদ্র্য ঘোচাতে ২০১৯ সালে শ্রমিক হিসেবে সৌদি আরব যান ফরিদপুরের ভাঙ্গা উপজেলার হাবিব খালাসী। কিন্তু গত বছর লাশ হয়ে দেশে ফেরেন তিনি। বিমান…	পরিবারের দারিদ্র্য ঘোচাতে ২০১৯ সালে শ্রমিক হিসেবে সৌদি আরব যান ফরিদপুরের ভাঙ্গা উপজেলার হাবিব খালাসী। কিন্তু গত বছর লাশ হয়ে দেশে ফেরেন তিনি।
international	সিডনিতে ছুরিকাঘাতে নিহতদের স্মরণে অপেরা হাউজে কালো ব্যাজ	অস্ট্রেলিয়ার সিডনির একটি শপিং সেন্টারে গণহত্যার শিকারদের শোক ও সম্মান জানাতে অস্ট্রেলিয়ার সবচেয়ে আইকনিক স্থাপত্যের ল্যান্ডমার্ক কালো ফিতা দিয়ে আলোকিত হয়েছে…	অস্ট্রেলিয়ার সিডনির একটি শপিং সেন্টারে গণহত্যার শিকারদের শোক ও সম্মান জানাতে অস্ট্রেলিয়ার সবচেয়ে আইকনিক স্থাপত্যের ল্যান্ডমার্ক কালো ফিতা দিয়ে আলোকিত হয়েছে।

Fig. 2. Sample Of Custom Dataset

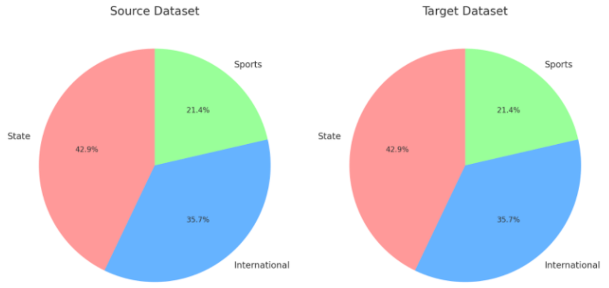


Fig. 3. Distribution of data for domain adaptation for source and target domain

21.4% for sports, as depicted in Fig. 3.

C. Transformer Model Architecture

The foundation of the proposed methodology lies in the transformer-based architectures, mT5 and flan-T5, which are particularly suited for sequence-to-sequence tasks like text summarization. These models leverage an encoder-decoder structure, ensuring robust handling of multilingual data and task-specific nuances. The process begins with the input sequence $X_i = \{x_1, x_2, \ldots, x_L\}$, representing Bangla text that has been preprocessed and tokenized. Each tokenized sequence is converted into embeddings $T_i \in \mathbb{R}^{L\times d}$, where L is the fixed sequence length and d represents the embedding dimension. Positional encodings $P_i \in \mathbb{R}^{L\times d}$ are added to these embeddings, resulting in $E_i = T_i + P_i$, which captures both semantic and positional information.

The encoder processes E_i through multiple layers of self-attention and feed-forward networks, generating latent representations $h_i = \{h_1, h_2, \ldots, h_L\}$, where each $h_i \in \mathbb{R}^m$ encapsulates the semantic and contextual features of the input sequence. These latent vectors are passed to the decoder, which uses cross-attention mechanisms to focus on relevant parts of h_i while generating the summary. The decoder predicts the output sequence $\hat{Y}_i = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{L'}\}$, where L' is the length of the summary. Each decoder output is transformed into a probability distribution over the vocabulary using a softmax function, ensuring accurate token prediction at each step.

The mT5 model, trained on a diverse multilingual dataset, enables the framework to handle Bangla text effectively by providing language-agnostic and context-aware representations. Flan-T5, fine-tuned on task-specific datasets, enhances summarization performance by leveraging instruction-tuned knowledge. Together, these models form a robust foundation for the summarization pipeline, which is further refined through domain adaptation mechanisms. Fine-tuning is performed using the XSum Bangla dataset and custom datasets specific to domains like state, international, and sports, ensuring adaptability to closed-set contexts.

The output probabilities for each token are computed as

$$P(\hat{y}_t|\hat{y}_{<t}, h_i) = \text{Softmax}(W_o z_t), \tag{1}$$

where z_t represents the decoder output at time step t, and W_o is the output projection matrix. To train the model, the cross-entropy loss for summary generation is defined as

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{L'} \log P(\hat{y}_t|\hat{y}_{<t}, h_i), \tag{2}$$

where N is the batch size. This loss ensures the generated summaries closely match the reference outputs. By combining these mechanisms with domain-specific fine-tuning, the proposed framework achieves robust and adaptable Bangla text summarization.

D. Training LLM models on Bengali Domain for Text Summarization

In this study, two Large Language Models - mT5 and flanT5 were used on the Bengali domain. For training the data, the bengali split of the XLSUM dataset from Hugging face containing 10.1k rows was acquired. Firstly, a tokenizer that will specifically work well on the Bengali language was developed.

After tokenization, the flanT5 model was trained on the train set of the bengali split of the XLSUM dataset, while the test split was kept as the unseen data for testing later on. Fig. 4 depicts the text and summary length distribution for both the train and test set. The flanT5 model and the trained custom tokenizer were loaded and special tokens like '<pad>' were added to the custom tokenizer and the model's token embedding was resized to match the custom tokenizer. The XLSUM dataset was loaded, and the 'text' and 'summary' columns were tokenized with the custom tokenizer and mapped using the map function of 'datasets' library completing the pre-processing of data. The flanT5 model was trained on 3 epochs with a learning rate of 5e-5, and attained training and evaluation loss of 1.5500 and 1.4400. The BLEU score and ROGUE score were obtained as 0.001 and 0.16 respectively for the model generated summary compared to the reference summary. Similarly, the mT5 model was also trained on the bengali split of the XLSUM dataset with an epoch of 10 and learning rate of 2e-5, procuring training loss of 3.5655 and validation loss of 1.1406. The model was further evaluated on the unseen test split of the XLSUM dataset achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.21, 0.09, and 0.20, respectively, alongside a BLEU score of 0.17 and BERTScore of 0.72.

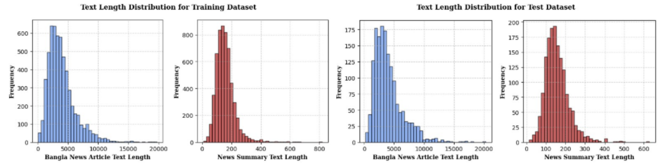


Fig. 4. Text length distribution for train and test dataset (XLSUM)

E. Domain Adaptation in Closed-Set Context

The proposed methodology applies domain adaptation within a closed-set context to address the gap between the source and target domains. In this context, the label spaces of both the source and target domains are identical, encompassing three specific categories: state, international, and sports. This setup ensures that the model focuses exclusively on learning invariant features across domains without introducing additional complexity due to label mismatch. The source domain dataset from Kaggle, BanglaMCT7, and the target domain dataset, derived via web scraping of Bengali newspaper articles, both include these three domains, providing a consistent label space for adaptation.

The closed-set domain adaptation process begins by generating latent representations h_i from the encoder, which captures semantic information from the input text. Since the label spaces are identical, the goal is to align the feature distributions of the source and target domains in this shared label space. This alignment is achieved through the domain classification and domain confusion losses. For a given domain label z_i, the domain discriminator computes probabilities $P_D(z_i|h_i)$, where h_i is the latent representation. The domain classification loss, expressed as:

$$L_D = -\frac{1}{N}\sum_{i=1}^{N}\big[z_i\log(\hat{z}_i) + (1-z_i)\log(1-\hat{z}_i)\big],$$

ensures that the discriminator correctly identifies whether a sample originates from the source or target domain.

In the closed-set context, it is crucial to minimize domain-specific features in the latent space h_i. This is achieved through the domain confusion loss, which leverages a gradient reversal layer (GRL). The GRL negates the gradient signal from the domain discriminator, encouraging the encoder to produce domain-invariant features. The domain confusion loss is defined as:

$$L_{DC} = -\frac{1}{N}\sum_{i=1}^{N}\log P_D(z_i|h_i),$$

where $P_D(z_i|h_i)$ measures the likelihood of h_i belonging to a specific domain. By minimizing L_{DC}, the encoder aligns the feature distributions of the source and target domains within the shared label space.

To address the summarization task, the latent features h_i are further processed by the decoder to generate summaries, optimized using the cross-entropy loss L_{CE}. The overall objective of the model is to minimize the total loss:

$$L_{\text{total}} = L_{CE} + \lambda_1 L_D + \lambda_2 L_{DC},$$

where λ_1 and λ_2 are hyperparameters controlling the contributions of domain adaptation losses. This total loss ensures that the model learns domain-invariant features while preserving the semantic consistency required for summarization.

The closed-set assumption simplifies the adaptation process by eliminating the need to map between different label spaces, allowing the model to focus on reducing the divergence between the source and target feature distributions. In the zero-shot setting, the model is evaluated on the target domain without further fine-tuning, testing its ability to generalize using the invariant features learned during training. In the fine-tuning setting, the model is further trained on the target domain to enhance its performance for domain-specific nuances. By working within the closed-set context, the methodology ensures effective domain adaptation without compromising the quality of generated summaries. Metrics such as ROUGE, BLEU, and BERTScore are utilized to evaluate the summarization performance across domains, demonstrating the robustness and versatility of the proposed approach.

F. Overall Objective Function

To optimize the performance of the domain-adaptive Bengali text summarization model, we employ a holistic approach that integrates domain adaptation-specific loss functions with the training dynamics of the mT5 architecture. The core objective is guided by the **cross-entropy loss** ($\mathcal{L}_{CE}$), which ensures effective learning of the summary generation task by penalizing discrepancies between the predicted tokens y_t and the ground truth tokens y_t^*. This is mathematically defined as:

$$\mathcal{L}_{CE} = -\frac{1}{T}\sum_{t=1}^{T}\log P(y_t^*|y_{<t}, x),$$

where x represents the input text, and T denotes the sequence length. To facilitate domain adaptation, we introduce a **domain classification loss** ($\mathcal{L}_D$), which utilizes domain labels $z \in \{0,1\}$. This binary cross-entropy loss encourages the model to distinguish between domains and is expressed as:

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N}\left[z_i\log\hat{z}_i + (1-z_i)\log(1-\hat{z}_i)\right],$$

where $\hat{z}_i$ represents the predicted domain label, and N is the number of samples in the batch. Furthermore, we employ a **domain confusion loss** ($\mathcal{L}_{DC}$), implemented via a gradient reversal layer to encourage domain-invariant latent representations. This is defined as:

$$\mathcal{L}_{DC} = -\frac{1}{N}\sum_{i=1}^{N}\log P_D(z_i|h_i),$$

where h_i is the latent representation of the i-th input. The final **total loss** combines these components with weighting factors λ_1 and λ_2 to balance the contributions of domain adaptation, given by:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_D + \lambda_2 \mathcal{L}_{DC}.$$

During the training of the mT5 model, the input sequence X and target summary Y are tokenized using a tokenizer T to ensure uniform sequence lengths L, as follows:

$$T(X) = \text{Pad}(\text{Truncate}(X, L), L).$$

The model is optimized using the AdamW optimizer with weight decay λ, and the learning rate is scheduled using cosine annealing with restarts:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{t\pi}{T_{\text{cur}}}\right)\right),$$

where T_{cur} represents the current cycle length. Additionally, gradient accumulation is employed to simulate a larger batch size:

$$\text{Effective Batch Size} = \text{Per Device Batch Size} \times G,$$

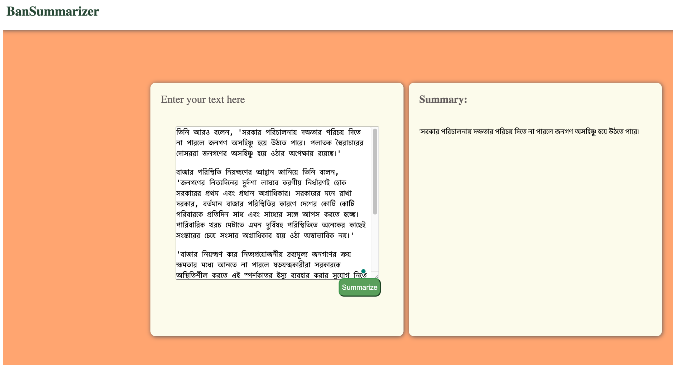
where G denotes the accumulation steps. These strategies, combined with domain-specific loss formulations, ensure robust domain adaptation and high-quality summarization performance.

G. Development of user Interface for the summary generation

For creating a user-friendly web interface for Bengali text summarization where users can input long Bengali text and receive a concise summary, web development frameworks like Flask were used for backend development, while frontend technologies such as HTML, CSS, and JavaScript were employed to design the interface. The backend handles requests from the frontend by loading models and performing text summarization. The backend receives the input text via the POST request method, the custom tokenizer from Hugging Face tokenizes the input text into token IDs, and the model generates a summary based on the input. Next, the backend sends output to the frontend as JSON, while HTML defines the structure of the page, input box, output text area, button for triggering summarization and CSS handles the styling of the input box, output area, button for a visually appealing interface. The fetch API of Javascript sends requests to the backend and Flask processes the request, runs the model and returns the result in JSON format. Finally, Javascript parses the JSON response and updates the DOM to display the output summary in the UI, thereby, aiding users to quickly and simply create insightful summaries from long, complex bengali texts. The web interface created in this study is depicted in Fig. 5.

IV. Investigation/Experiment, Result, Analysis and Discussion

This section elaborates on the results that are obtained from the proposed study - bengali text summarization



তিনি আরও বলেন, 'সরকার পরিচালনায় দক্ষতার পরিচয় দিতে না পারলে জনগণ অসহিষ্ণু হয়ে উঠতে পারে। পলাতক স্বৈরাচারের দোসররা জনগণের অসহিষ্ণু হয়ে ওঠার অপেক্ষায় রয়েছে।'

বাজার পরিস্থিতি নিয়ন্ত্রণের আহ্বান জানিয়ে তিনি বলেন, 'জনগণের নিত্যদিনের দুর্দশা লাঘবে করণীয় নির্ধারণই হোক সরকারের প্রথম এবং প্রধান অগ্রাধিকার। সরকারের মনে রাখা দরকার, বর্তমান বাজার পরিস্থিতির কারণে দেশের কোটি কোটি পরিবারকে প্রতিদিন সাধ এবং সাধ্যের সঙ্গে আপস করতে হচ্ছে। পারিবারিক খরচ মেটাতে এমন দুর্বিষহ পরিস্থিতিতে অনেকের কাছেই সংস্কারের চেয়ে সংসার অগ্রাধিকার হয়ে ওঠা অস্বাভাবিক নয়।'

'বাজার নিয়ন্ত্রণ করে নিত্যপ্রয়োজনীয় দ্রব্যমূল্য জনগণের ক্রয় ক্ষমতার মধ্যে আনতে না পারলে ষড়যন্ত্রকারীরা সরকারকে অস্থিতিশীল করতে এই স্পর্শকাতর ইস্যু ব্যবহার করার সুযোগ নিতে

'সরকার পরিচালনায় দক্ষতার পরিচয় দিতে না পারলে জনগণ অসহিষ্ণু হয়ে উঠতে পারে।

Fig. 5. Web interface of the project

leveraging large language models and domain adaptation on closed sets. Here, ROGUE score, BLEU score and BERTSCORE were used as evaluation parameters to assess the performance of the different LLM models for both bengali text summarization and domain adaptation on fine-tuning and zero-shot settings.

A. Training Metrics and Summary for Bengali Text Summarization

Table II presents a comprehensive training summary including hyperparameters and training outcomes of the two LLM models - flant5 and mt5 which are used in this study and are employed in the summarization of the bengali texts in the dataset. The training process involved fine-tuning these models on the Bengali corpus to evaluate their performance in generating summaries.

TABLE II
TRAINING SUMMARY OF THE MODELS

Models	**Learning Rate**	**Epoch**	**Training Loss**	**Validation Loss**
flanT5	5e-5	3	1.5500	1.4400
mT5	2e-5	10	3.5655	1.1406

Both the LLM models - flant5 and mt5 leveraged in this study were trained initially on the bengali split of the XL-SUM dataset. The mt5 model was trained on an epoch of 10 with a learning rate of 2e-5. The model being trained for 10 epochs ensured sufficient exposure to the training data for convergence without overfitting, while the choice of the learning rate as 2e-5 was optimal for minimizing overfitting and achieving convergence within the 10 epochs. The training loss and validation loss were reported to be 3.5655 and 1.1406 respectively. This training loss is generally acceptable for complex tasks like text summarization, while the validation loss being slightly lower than the training loss indicates effective generalization, suggesting that the model performs well on unseen data. Similarly, the flant5 model was trained on an epoch of 3 with a learning rate of 5e-5. This relatively high learning rate suits the smaller number of epochs, providing faster updates to the model

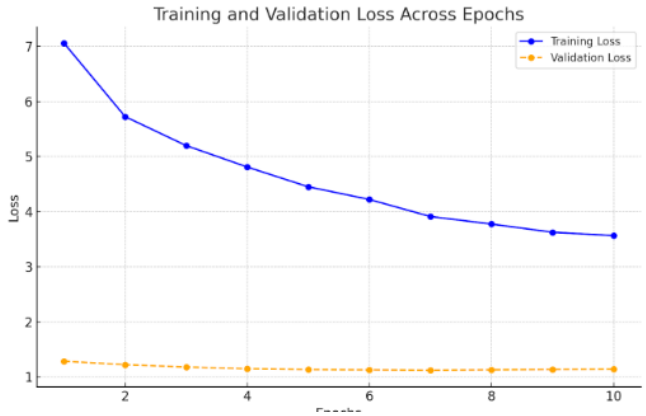


Fig. 6. Training metrics for mT5 model

parameters. Also, the gap between the training loss and validation loss further indicates a well-regularized model. The training metrics for the models are also displayed in a graph as shown in Fig. 6. Despite both LLM models showing strong generalization capabilities, however, the mT5 model's lower validation loss suggests it is more robust in adapting to the complexities of a low-resource language like Bengali.

B. Evaluation Parameters of the Models for Bengali Text Summarization

The evaluation parameters of the two LLM models after training were examined as demonstrated in Table III, providing important context about the diagnostic capacities of the models. The models were assessed using three evaluation metrics: ROUGE, BLEU, and BERTScore.

ROUGE is a set of metrics used to evaluate the quality of summaries or translations by comparing them to reference summaries or translations. It measures the overlap of n-grams (contiguous sequences of n items, typically words) between the generated summary and the reference summaries. The type of ROUGE metrics used for this research are -

- ROUGE-N: This measures the overlap of n-grams between the generated summary and reference summaries. It includes ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-3 (trigrams), etc.
- ROUGE-L: This measures the longest common subsequence (LCS) between the generated summary and reference summaries.

BLEU is a metric for evaluating the quality of machine-translated text by comparing it to one or more reference translations. It measures the precision of n-grams (contiguous sequences of n items, typically words) in the generated translation compared to the reference translations. BLEU computes a precision score for each n-gram up to a certain length (usually up to 4 grams) and combines them using a weighted geometric mean.

BERTScore uses BERT embeddings to evaluate the similarity between the generated summary and the reference summary at a semantic level, rather than just surface-level word overlap. It calculates precision, recall, and F1-score by computing cosine similarities between BERT embeddings of words in the generated and reference summaries.

TABLE III
EVALUATION PARAMETERS OF THE MODELS FOR BENGALI TEXT SUMMARIZATION

Models	**Evaluation Parameter Scores**			**BLEU**	**BERT**
	ROUGE-1	**ROUGE-2**	**ROUGE-L**		
flanT5	0.15	0.04	0.10	0.02	0.76
mT5	0.24	0.09	0.20	0.17	0.72

Following the training process, the performance evaluation of flanT5 and mT5 models were carried out on the test split of the XLSUM dataset that was the unseen data during training. The models were assessed using the three evaluation metrics as defined above: ROUGE, BLEU, and BERTScore. The weighted ROUGE F1-measure scores are utilized as the primary metric to compare the performance of the models. The mt5 model achieved the best ROUGE scores among the two LLM models on the XLSum dataset, where it achieved ROUGE-1 score of 0.21, ROUGE-2 score of 0.09 and ROGUE-L score of 0.20. Further results revealed that the mt5 model outperformed the flant5 model across all the metrics, also achieving an impressive BLEU score and BERTSCORE of 0.17 and 0.72 respectively. However, flanT5, while performing comparatively lower, achieved a BERTScore of 0.60, suggesting a moderate level of semantic similarity with the reference summaries. These evaluation parameter results indicate that mT5, with its multilingual training base, is better suited for Bengali text summarization tasks in this setting. Table IV further illustrates the summaries generated by the two LLM models compared to the reference summary of an example text.

C. Evaluation Parameters for Domain Adaptation in different settings

For domain adaptation, the two LLM models mT5 and flanT5 previously trained on the XLSUM Bangla corpus and then on XLSUM (Bengali) dataset for text summarization were used. The BanglaMCT7 dataset was used as the source dataset and the custom dataset was used as the target dataset. Under the closed-set domain adaptation technique, three domains- state, international, and sports were chosen from both the BanglaMCT7 and custom dataset and were then merged separately to be named as the source and target domains respectively. The models adapted to the three specific domains were tested in fine-tuning and zero-shot settings. Table V provides a detailed overview of the evaluation parameters (ROUGE, BLEU, BERT) obtained in these specific settings for the two different models.

For the fine-tuning setting, the model was trained on the source domain which is the merged dataset of the 3 specific domains - state, international, sports from the BanglaMCT7 dataset, while the target domain is also the merged dataset of the 3 domains from the custom dataset.

TABLE IV
EXAMPLE OF GENERATED SUMMARIES USING THE TWO MODELS

Section	**Content**
Article	ভারতের অন্য অঞ্চলেও কোক, পেপসি নিষিদ্ধ করার দাবি জানাচ্ছেন কর্মীরা। স্থানীয় পণ্যের ব্যবহার নিশ্চিত করার জন্যই এই উদ্যোগ গ্রহণ করেছে ব্যবসায়ীরা। রাজ্যের শীর্ষ দুটি ব্যবসায়ী এসোসিয়েশন এই দুটি পানীয় নিষিদ্ধ করার প্রস্তাব করেছিল। তারই প্রেক্ষাপটে আজ বুধবার থেকে তামিলনাডু রাজ্যে নিষিদ্ধ হলো কোকা-কোলা ও পেপসি। প্রতিষ্ঠানগুলো বলছে, কোমল পানীয়ের প্রতিষ্ঠানগুলো নদী থেকে প্রচুর পানি ব্যবহার করে, সেকারণে কৃষকদের জমি সেচের সময়ও ব্যাপক ভোগান্তিতে পড়তে হয়। বিশেষ করে খরার সময় সেচে পানি সমস্যা প্রকট হয়ে দাঁড়ায়। রাজ্যের দশ লাখেরও বেশি দোকানদার এ নিষেধাজ্ঞা মেনে চলবে বলে ধারণা করা হচ্ছে। গত মাসে তামিলনাডুতে 'জাল্লিকাটু' নামে ঐতিহ্যবাহী ষাঁড়ের লড়াই নিষিদ্ধের বিরুদ্ধে ব্যাপক বিক্ষোভের ঘটনা দেখে রাজ্যে পেপসি, কোকা-কোলা নিষিদ্ধের প্রস্তাব করে শীর্ষ দুটি ব্যবসায়ী সংগঠন ফেডারেশন অব তামিলনাডু ট্রেডার্স এসোসিয়েশন (এফটিএনটিএ) এবং তামিলনাডু ট্রেডার্স এসোসিয়েশন। বিক্ষোভের সময় অনেকে বলছিলেন 'জাল্লিকাটু' নিষিদ্ধ করা মানে স্থানীয় ঐতিহ্য ও সংস্কৃতিকে অবমাননা করা। "আমরা কয়েক মাস আগে কোমল পানীয়ের বিরুদ্ধে আমাদের প্রচারণা শুরু করি, কিন্তু যখন আমরা 'জাল্লিকাটু' নিষিদ্ধের প্রতিবাদে বিক্ষোভ শুরু করি, কোমল পানীয়ের বিরুদ্ধে আমাদের প্রচারণাও ভিন্ন রূপ পায়"- বিবিসি তামিল সার্ভিসকে দেয়া এক সাক্ষাৎকারে বলছিলেন এফটিএনটিএ'র প্রেসিডেন্ট থা ভেলায়ান। "পেপসি এবং কোকা-কোলার মতো পানীয় কিন্তু আপনার স্বাস্থ্যের জন্য ভালো নয়। কারণ এর মধ্যে বিভিন্ন ধরনের কেমিকেল থাকে এবং অতিরিক্ত চিনি থাকে এসব পানীয়তে। আমরা ভারতীয় কোমল পানীয়ের প্রচার চালাচ্ছি এবং ফলের জুসের বিক্রি যেন আরও বাড়ে সেই চেষ্টাও আমরা চালাবো"-বলছিলেন ব্যবসায়ী থা ভেলায়ান। স্থানীয় ব্যবসা এবং কৃষকদের উন্নতির কথা ভেবে সুপারমার্কেট, রেস্টুরেন্ট ও হোটেলগুলো যেন এই নিষেধাজ্ঞা মেনে চলে সেই আহ্বানও জানিয়েছে এসোসিয়েশনগুলো। এই নিষেধাজ্ঞার বিষয়ে পেপসি ও কোকা-কোলা প্রতিষ্ঠানের পক্ষ থেকে এখনও কোনো মন্তব্য পাওয়া যায়নি।
Reference Summary	ভারতের দক্ষিণাঞ্চলীয় রাজ্য তামিলনাডুর ব্যবসায়ীরা সেখানে কোকা-কোলা ও পেপসি বিক্রি নিষিদ্ধ ঘোষণা করেছে।
Model (mT5)	ভারতের তামিলনাডু রাজ্যে কোকা-কোলা এবং পেপসি নিষিদ্ধ করেছে ব্যবসায়ীরা। এই দুটি পানীয় নদী থেকে প্রচুর পানি ব্যবহার করে, সে কারণে কৃষকদের জমি সেচের সময়ও ব্যাপক ভোগান্তিতে পড়ে।
Model (flanT5)	ভারতের অন্য ব্যবহার নিষিদ্ধ করার জন্যে এই উদ্যোগ গ্রহণ করেছে ব্যবসায়ীরা।

TABLE V
EVALUATION PARAMETERS FOR 3 DIFFERENT DOMAINS LEVERAGING LLM MODELS IN DIFFERENT SETTINGS

Models	**State**			**International**			**Sports**		
	ROUGE	**BLEU**	**BERT**	**ROUGE**	**BLEU**	**BERT**	**ROUGE**	**BLEU**	**BERT**
Fine-Tuning Setting									
flanT5	0.79	0.13	0.86	0.78	0.12	0.85	0.79	0.11	0.86
mT5	0.41	0.31	0.78	0.44	0.36	0.83	0.35	0.27	0.81
Zero-Shot Setting									
flanT5	0.57	0.38	0.85	0.67	0.47	0.89	0.59	0.34	0.88
mT5	0.65	0.47	0.84	0.45	0.38	0.84	0.33	0.23	0.82

The two LLM base models were trained on the source domain, and then fine-tuned over the target domain, to achieve the ROUGE, BLEU and BERT scores separately for the 3 domains. As seen from the results achieved, the flant5 model outperformed the mt5 model for ROUGE and BERT scores, while the mt5 model performed better in the BLEU score evaluation. The state domain performed better than the other domains for the fine-tuning setting where for the flant5 model the ROUGE score is 0.79, BLEU score is 0.13 and BERTscore is 0.86, thereby, exhibiting a superior performance.

Next, zero-shot setting was performed on the models adapted to the three specific domains as before, where the domain specific models were tested in an unseen domain specific dataset of 500 rows for state, international and sports categories separately, that were collected individually by scraping online bengali news portals. For evaluating the results, only inference was carried out as this is the criteria for zero-shot setting, and the models were assessed using ROUGE, BLEU and BERT scores for the three separate domains. Upon evaluation, it can be seen that the mt5 model performed better for the state domain, while the flant5 model accomplished better results for the two other domains. Overall, the highest values were obtained for the international domain, where for the flant5 model, the ROUGE score is 0.67, BLEU score is 0.47 and BERT score is 0.89. All the models highlighted competitive scores therefore showcasing their capability to perform well under zero-shot conditions.

D. Discussion of the Result Analysis

This study aimed to explore the performance of two LLM models, flanT5 and mt5, across multiple configura-

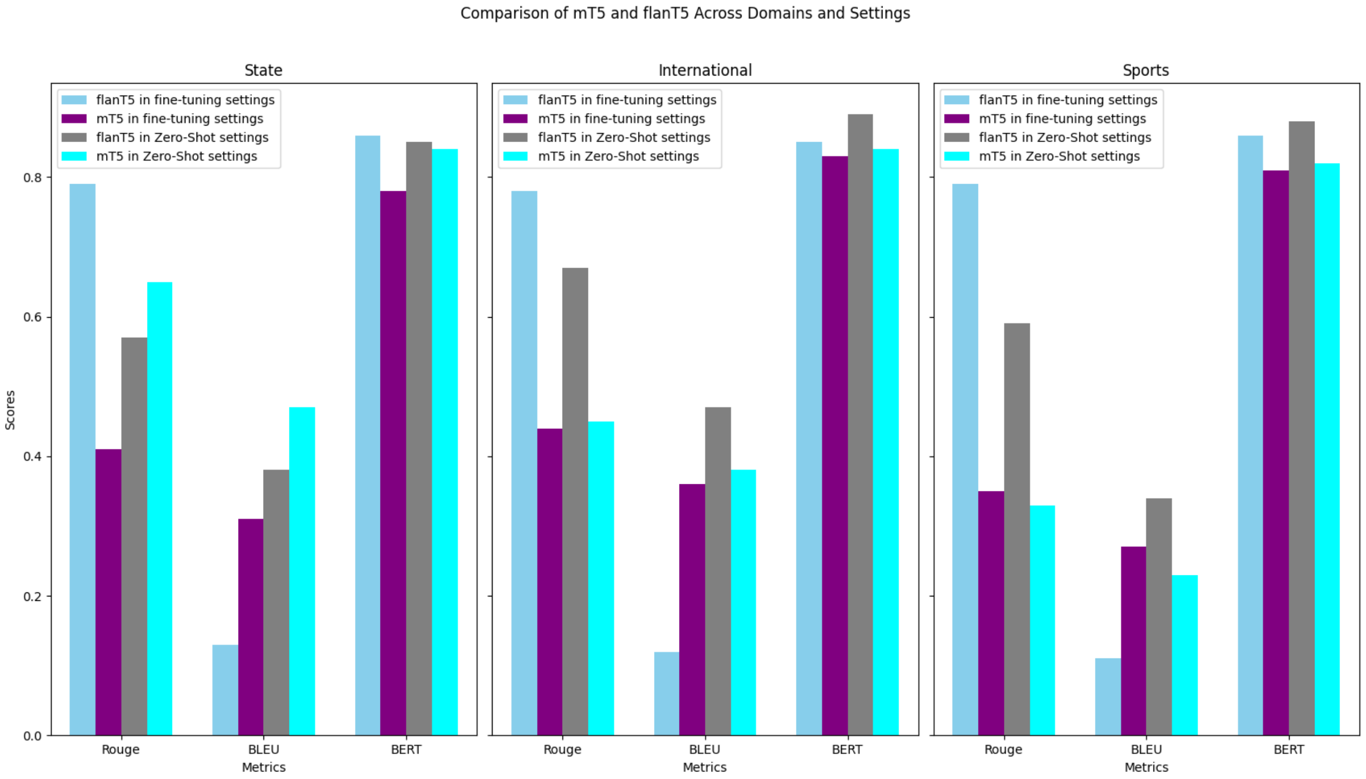


Fig. 7. Comparison of evaluation parameters of mT5 and flanT5 model across 3 domains in different settings

TABLE VI
A Comparison Of Existing Bangla Text Summarization Models With Our Approach

Study	Model/Approach	Dataset	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Observations
Bhattacharjee et al. (2020)	Seq2Seq with Attention	BANSData (19,096 samples)	0.30	—	0.31	0.30	Challenges with long inputs; issues with repetition and factual inaccuracies.
Dhar et al. (2021)	Hybrid Pointer Generator Network	BANSData & BANS-133k	0.67	0.42	0.41	—	Improved accuracy; reduced repetition; stable on large datasets.
Shahariar et al. (2023)	Ranking-Based Approach	Bangla Text Summarization	0.45	0.18	0.41	0.02	Enhanced summary accuracy by combining model outputs.
Our Approach (2025)	Domain-Adaptive FlanT5	Custom Dataset	0.67	0.57	0.67	0.46	Demonstrated superior performance in domain adaptation tasks.

tions, evaluation parameters, and domain-specific scenarios. Three key evaluation metrics - ROUGE, BLEU, and BERTScore, were utilized to assess the performance of the models under fine-tuning and zero-shot settings. The growing need for effective natural language processing (NLP) tools in low-resource languages like Bengali served as the motivation behind this project, as despite being spoken by millions of people around the world, the Bengali language still lacks robust tools for text summarization and domain adaptation compared to widely spoken languages. Therefore, our project contributes to the advancement of NLP for the Bengali language, promoting improved usability and accessibility in linguistic information processing.

For bengali text summarization, the two LLM models were trained on the bengali split of the XLSUM dataset, where both the models were trained on different hyperparameters and number of epochs. Following this, evaluation metrics were determined for the two models using ROUGE, BLEU and BERTScore which highlighted the strengths and weaknesses of the two models. The mt5 model outperformed the flanT5 model across most evaluation metrics, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.21, 0.09, and 0.20, respectively, alongside a BLEU score of 0.17 and BERTScore of 0.72. These metrics highlight the effectiveness with which mT5 performs on complex Bengali text summarization tasks, which can be ascribed to its multilingual pretraining on a variety of datasets, thereby, allowing it to achieve competitive results, particularly in resource-constrained settings. Table VI compares the performance of various

Bengali text summarization models based on ROUGE and BLEU metrics from past notable research publications. The comparison highlights the improvements achieved by the proposed domain-adaptive FlanT5 model, which outperforms other approaches in domain adaptation tasks, achieving superior scores across all metrics.

Additional information on the models' domain-specific adaptability was obtained from the domain adaptation tests, which were carried out in fine-tuning and zero-shot configurations. The performance of the state, international, and sports datasets was assessed using ROUGE, BLEU, and BERTScore measures after they were meticulously selected. In the fine-tuning setting, flanT5 performed better in ROUGE and BERTScore metrics, and flanT5 demonstrated its strong generalization and semantic alignment abilities in the state domain with a ROUGE score of 0.79, BLEU score of 0.13, and BERTScore of 0.86. On the other hand, mT5 demonstrated superior accuracy in n-gram matching, as evidenced by its higher BLEU scores. In the zero-shot setting, the models were evaluated on unseen domain-specific datasets without additional training, and the outcomes showed that both models performed competitively. Notably, the highest performance was observed for the international domain in the zero-shot setting, where flanT5 achieved a ROUGE score of 0.67, BLEU score of 0.47, and BERTScore of 0.89. Fig. 7 further illustrates the comparison of evaluation parameters of mT5 and flanT5 model across the three different domains in different settings. These findings illustrate the strong zero-shot generalization ability of flanT5, especially in linguistically rich and diverse environments.

As the last step after achieving all the training and evaluation parameter results, a user-friendly web application was developed using the Flask framework of Python. This application serves as an accessible interface for Bengali text summarization, where on the left, users input Bengali text into the box, and upon clicking the ”Summarize” button, the summary is generated and displayed in the ”Summary” section on the right. This interface aids users to rapidly and effectively get concise summaries of long Bengali texts. The user-friendly design of the interface makes it possible for people without technical knowledge to quickly and simply create insightful summaries, which makes it appropriate for a variety of uses like document analysis, educational content, and news summarizing.

This study demonstrates the potential of leveraging LLMs for Bengali text summarization and domain adaptation. By integrating powerful LLM models and domain adaptation techniques, the project contributes to acquiring accurate Bengali text summarization tools that can be accessed by all.

V. Impacts of the Project

A. Impact of this project on societal, health, safety, legal and cultural issues

The enhanced access to critical information for over 250 million Bengali speakers can be attained by summarizing texts in Bengali. It has the potential to be of great use for individuals with busy schedules and can also significantly facilitate slow readers. In the educational sector, summarized Bengali texts can help students and teachers by providing shortened versions of complicated materials which is especially convenient for English medium background students.

Medical documents or reports in Bengali can be summarized making healthcare information more attainable and easier to read. During healthcare emergencies, the domain-adapted summaries could simplify the distribution of instructions, helping save lives through rapid and clear communication.

During periods of crisis or natural disasters, summarized Bengali disaster management protocols, weather reports or government notices could improve public safety in those difficult times. Additionally, its safety and caution could also be guaranteed as it reduces the risk of misunderstanding by making things simpler.

In rural communities, citizens are comparatively more oblivious to their legal rights and obligations therefore it will enable them to understand it more clearly. By exploiting LLMs for Bengali, the project enhances and encourages the use of this low-resource language in the digital space which as a result sustains its cultural heritage. Thus it shows that every individual, regardless of their language, can benefit from technology advancements.

B. Impact of this project on environment and sustainability

The issues of deforestation and lower carbon emissions from paper production can be restrained by reducing the need for printing lengthy documents which in turn saves paper. This project encourages users for digital solutions instead of relying on printed materials such as books, reports, or lengthy documents. Hence, the transition from traditional methods to digital resources not only makes information accessible to a wider audience but also plays an important role in environmental sustainability. Leveraging LLMs in this project ensures minimum human efforts and energy consumption compared to manual methods and it provides long-term sustainable practices in data handling. The project's focus on domain adaptation and that too on a low-resource language like Bengali can encourage sustainable approaches for other low-resource languages promoting equality in access to technology and information worldwide.

VI. Project Planning

Fig. 8 depicts the project planning through a gantt chart for the duration of CSE499A and CSE499B. Each group member's contribution and the total work timeline is highlighted in the figure below.

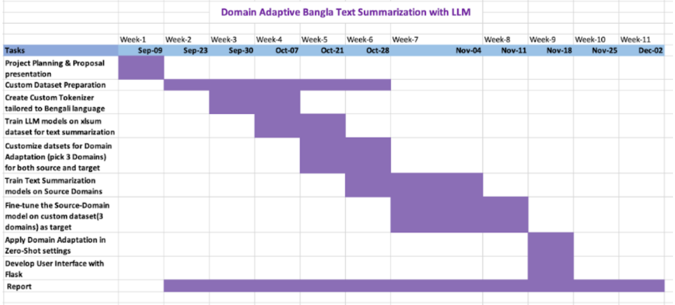


Fig. 8. Gantt chart for the project

VII. Conclusion and Future Works

A. Summary

Overall, this project has experimented with different types of LLMs for Bangla text summarization, out of which the Flant5 model and the mt5 models were actually able to generate summaries. However, mt5 model was able to generate better summaries. These models were trained on the XLSum dataset from huggingface. After this we have created our own custom dataset which contains 21,512 rows. This custom dataset was later categorized across seven different categories, out of which the international and state category had the highest frequency. This was done as categorizing basically creates domains for Domain Adaptation to be applied. The models were fine-tuned on three domains which were state, international and sports. The models were trained and assessed under fine-tuning and zero-shot conditions, using ROUGE, BLEU, and BERTScore metrics to evaluate performance. The FlanT5 model performed exceptionally well in fine-tuning, especially in the state domain, achieving ROUGE and BERTScore metrics as high as 0.79 and 0.86, respectively. In contrast, the mT5 model showed competitive results in zero-shot scenarios, highlighting its ability to adapt to unseen data. Notably, the international domain recorded the best performance metrics in zero-shot settings with FlanT5, achieving a ROUGE score of 0.67, a BLEU score of 0.47, and a BERTScore of 0.89.

B. Limitations

The custom dataset is limited to seven categories, with only three (state, international, and sports) utilized for domain adaptation. This limitation affects the models' ability to generalize across a wider range of domains. The evaluation process mainly depends on automated metrics such as ROUGE, BLEU, and BERTScore, which may not adequately reflect the cultural and linguistic subtleties of Bengali text. Training and fine-tuning large language models like mT5 and flanT5 demand considerable computational resources, which may not be practical in environments with limited resources. Although zero-shot evaluations showed competitive results, their performance varied by domain, suggesting that the models might have difficulty with specific nuances in unfamiliar contexts.

C. Future Improvement

Future directions include expanding the dataset to cover additional domains such as healthcare, technology, and legal documents to improve the model's adaptability and application range. Conducting qualitative evaluations alongside automated metrics would yield deeper insights into the cultural and contextual accuracy of the summaries. Testing cutting-edge models like GPT-4 or BongLLaMA could provide better benchmarks and insights into Bengali text summarization. Improving the user interface to facilitate real-time feedback and a more intuitive design would enhance accessibility and user experience. Additionally, exploring adaptation techniques such as partial and universal domain adaptation could lead to a more thorough understanding of model performance across various settings. Data augmentation strategies should also be employed.

References

[1] L. Abualigah, M. Q. Bashabsheh, H. Alabool, and M. Shehab, "Text summarization: a brief review," *Recent Advances in NLP: the case of Arabic language*, pp. 1–15, 2020.

[2] M. S. Islam, "Research on bangla language processing in bangladesh: progress and challenges," in *8th international language & development conference*, 2009, pp. 23–25.

[3] N. Dhar, G. Saha, P. Bhattacharjee, A. Mallick, and M. S. Islam, "Pointer over attention: An improved bangla text summarization approach using hybrid pointer generator network," in *2021 24th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2021, pp. 1–5.

[4] G. Shahariar, T. Talukder, R. A. K. Sotez, and M. T. R. Shawon, "Rank your summaries: Enhancing bengali text summarization via ranking-based approach," in *International Conference on Big Data, IoT and Machine Learning*. Springer, 2023, pp. 153–167.

[5] P. Bhattacharjee, A. Mallick, and M. Saiful Islam, "Bengali abstractive news summarization (bans): a neural attention approach," in *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*. Springer, 2021, pp. 41–51.

[6] G. Sharma and D. Sharma, "Automatic text summarization methods: A comprehensive review," *SN Computer Science*, vol. 4, no. 1, p. 33, 2022.

[7] A. Afzal, R. Chalumattu, F. Matthes, and L. Mascarell, "Adapteval: Evaluating large language models on domain adaptation for text summarization," *arXiv preprint arXiv:2407.11591*, 2024.

[8] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, "A survey on domain adaptation theory: learning bounds and theoretical guarantees," *arXiv preprint arXiv:2004.11829*, 2020.

[9] Y. Li, S. Miao, H. Huang, and Y. Gao, "Word matters: What influences domain adaptation in summarization?" *arXiv preprint arXiv:2406.14828*, 2024.

[10] H. Zhang, P. S. Yu, and J. Zhang, "A systematic survey of text summarization: From statistical methods to large language models," *arXiv preprint arXiv:2406.11289*, 2024.

[11] Z. Zhao and P. Chen, "To adapt or to fine-tune: A case study on abstractive summarization," in *China National Conference on Chinese Computational Linguistics*. Springer, 2022, pp. 133–146.

[12] T. Yu, Z. Liu, and P. Fung, "Adaptsum: Towards low-resource domain adaptation for abstractive summarization," *arXiv preprint arXiv:2103.11332*, 2021.

[13] A. Afzal, J. Vladika, D. Braun, and F. Matthes, "Challenges in domain-specific abstractive summarization and how to overcome them," *arXiv preprint arXiv:2307.00963*, 2023.

[14] L. Basyal and M. Sanghvi, "Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models," *arXiv preprint arXiv:2310.10449*, 2023.

[15] M. Borah, P. Dadure, P. Pakray *et al.*, "Comparative analysis of t5 model for abstractive text summarization on different datasets," 2022.

[16] M. Lewis, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.

[17] H. Song, H. Su, I. Shalyminov, J. Cai, and S. Mansour, “Finesure: Fine-grained summarization evaluation using llms,” *arXiv preprint arXiv:2407.00908*, 2024.

[18] Z. Kolagar and A. Zarcone, “Aligning uncertainty: Leveraging llms to analyze uncertainty transfer in text summarization,” in *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, 2024, pp. 41–61.

[19] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, “Bangla-bert: transformer-based efficient model for transfer learning and language understanding,” *IEEE Access*, vol. 10, pp. 91 855–91 870, 2022.

[20] R. R. Chowdhury, M. T. Nayeem, T. T. Mim, M. S. R. Chowdhury, and T. Jannat, “Unsupervised abstractive summarization of bengali text documents,” *arXiv preprint arXiv:2102.04490*, 2021.

[21] B. Jahan, S. S. Mahtab, M. Faizul Huq Arif, I. S. Emon, S. A. Milu, and M. Julfiker Raju, “An automated bengali text summarization technique using lexicon-based approach,” in *Innovations in Computer Science and Engineering: Proceedings of 8th ICICSE*. Springer, 2021, pp. 363–373.

[22] M. A. T. Rony and M. S. Islam, “Evaluating large language models for summarizing bangla texts,” in *Eighth Widening NLP Workshop (WiNLP 2024) Phase II*.

[23] S. K. Lora and R. Shahriyar, “Conversum: A contrastive learning based approach for data-scarce solution of cross-lingual summarization beyond direct equivalents,” *arXiv preprint arXiv:2408.09273*, 2024.

[24] J. Parnell, I. J. Unanue, and M. Piccardi, “Sumtra: A differentiable pipeline for few-shot cross-lingual summarization,” *arXiv preprint arXiv:2403.13240*, 2024.

[25] A. K. Zehady, S. A. Mamun, N. Islam, and S. Karmaker, “Bongllama: Llama for bangla language,” *arXiv preprint arXiv:2410.21200*, 2024.

[26] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, “A brief review of domain adaptation,” *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[27] P. P. Busto, A. Iqbal, and J. Gall, “Open set domain adaptation for image and action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 413–429, 2018.

[28] Z. Cao, M. Long, J. Wang, and M. I. Jordan, “Partial transfer learning with selective adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2724–2732.

[29] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Universal domain adaptation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2720–2729.

[30] W. Lu, R. K. Luu, and M. J. Buehler, “Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities,” *arXiv preprint arXiv:2409.03444*, 2024.

[31] J. Zhong and Z. Wang, “Mtl-das: Automatic text summarization for domain adaptation,” *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 4851828, 2022.

[32] Y. Mao, M. Zhong, and J. Han, “Citesum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision,” *arXiv preprint arXiv:2205.06207*, 2022.