Abstractive Summarization of Low resourced Nepali language using Multilingual Transformers

Prakash Dhakal[1], **Daya Sagar Baral**[1]
[1]Department of Electronics and Computer Engineering,
Institute of Engineering, Pulchowk Campus,
Tribhuvan University, Nepal
Correspondence: 079msdsa016.prakash@pcampus.edu.np

Abstract

Nepali, one of the prominent languages of South Asia, remains underrepresented in natural language processing (NLP) research, particularly in the domain of abstractive summarization. While significant progress has been made in extractive summarization, the complexity of generating coherent, human-like summaries from low-resource languages like Nepali is still largely unexplored. This paper introduces the first comprehensive study on applying multilingual transformer-based models, specifically mBART and mT5, to the task of generating headlines for Nepali news articles through abstractive summarization. To address the absence of large-scale datasets for this task, we developed a Nepali news headline summarization corpus by aggregating data from multiple online news portals. Leveraging this dataset, we fine-tuned multilingual transformer models, mBART and mT5, using Low-Rank Adaptation (LoRA) and quantization techniques to optimize computational efficiency without sacrificing performance. Comprehensive evaluations were conducted using ROUGE scores to measure the models' output quality, complemented by a detailed human evaluation to select the best summary overall based on relevance, fluency, conciseness, informativeness, factual accuracy, and coverage. Notably, the 4-bit quantized mBART model demonstrated superior performance, significantly reducing computational costs while maintaining high-quality results. This work not only underscores the feasibility of applying transformer-based approaches to Nepali abstractive summarization but also provides a scalable solution to advancing NLP capabilities for underrepresented South Asian languages.

Keywords: Nepali Abstractive text summarization, Transformers, Natural language processing, Low-Rank Qdaptation (LoRA), Quantization, ROUGE, Human evaluation

1 Introduction

The exponential growth of digital content, such as news articles, blogs, and social media, has made automatic text summarization a critical task in Natural Language Processing (NLP). This involves generating concise summaries that capture the main ideas of the original text while maintaining its meaning. Summarization is generally performed in two ways: extractive summarization and abstractive summarization. Abstractive summarization generates new sentences to convey the original text's meaning, requiring sophisticated language generation, while extractive summarization involves the extraction of key sentences or phrases from the original text.

Summarization in Nepali language plays a crucial social and practical role, particularly in areas such as education, news aggregation, and information access. In rural communities and underserved populations, where internet infrastructure is limited, concise and relevant summaries can help bridge the information gap. Additionally, in the context of education, this technology can generate brief and informative content summaries to aid students and educators. This research not only contributes to enhancing the digital content accessibility for Nepali speakers but also highlights the potential for large-scale deployment in sectors that rely heavily on information dissemination, making it highly relevant to the region's linguistic needs.

Transformer models such as mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) have proven to be highly effective for a variety of NLP tasks in low-resource languages, offering state-of-the-art performance in text generation and summarization tasks. These models utilize the transformer architecture, which is adept at capturing long-range dependencies in text, making them particularly suitable for abstractive summarization where the model must generate coherent, novel sentences rather than merely extracting phrases from the

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 124–133
January 19, 2025.

source text. Compared to earlier approaches that relied on recurrent neural networks (RNNs) like GRU (Gated Recurrent Units) or LSTMs (Long Short-Term Memory networks), transformers are able to process input sequences in parallel, making them more efficient and scalable for large datasets.

This study represents the first known application of transformer models, specifically mBART(Liu et al., 2020) and mT5(Xue et al., 2021), for abstractive summarization in the Nepali language. It introduces a novel dataset, and by leveraging LoRA with quantization techniques to optimize performance for low-resource settings. This research marks a crucial step forward for underrepresented Nepali languages in NLP. A novel Nepali news summarization dataset had to be created by scraping data from various news portals due to lack of dataset for this particular task. The multilingual models were then fine-tuned with this dataset using Low-Rank Adaptation (Hu et al., 2021) and quantization techniques as suggested in (Dettmers et al., 2023), making the training process more computationally efficient and faster. The performance of these models were then evaluated using ROUGE scores (Lin, 2004) and human evaluation following winner-take-all approach based on criteria such as relevance, fluency, conciseness, informativeness, factual accuracy, and coverage to ensure the generated summaries were coherent and conveyed the original meaning.

2 Related Work

With the rise of transformer-based models (Vaswani et al., 2023), various research works have been carried out using them for text summarization. Many studies focus on English, while research on the Nepali language is limited and primarily based on extractive summarization approaches.

(Ranabhat et al., 2019) introduced extractive summarization to produce summaries from multiple Nepali sentences by selecting a subset from the original text using TextRank (Mihalcea and Tarau, 2004). These summaries contained the most important sentences of the input. They utilized TextRank for sentence scoring and topic modeling for summary evaluation.

(Mishra et al., 2020) generated Nepali news headlines using GRU (Chung et al., 2014) in an encoder-decoder fashion, taking the news content as input and generating a headline as output. The news was converted into word tokens and vectorized using FastText (Bojanowski et al., 2017), trained on a corpus of Nepali news articles and headlines collected from several web portals.

(Khanal et al., 2022) employed an extractive method for Nepali text summarization using TextRanking (Mihalcea and Tarau, 2004) and LSTM (Hochreiter and Schmidhuber, 1997). They trained a Nepali news corpus with GloVe embeddings using different window sizes (10, 12, 15) and vector sizes (100, 200, 300). For extractive text summarization, they used Text Ranking and an attention-based LSTM model (Wang et al., 2016).

(Timalsina et al., 2022) introduced an attention-based RNN for abstractive Nepali text summarization. They first created a Nepali text dataset by scraping Nepali news from online portals, then designed a deep learning-based summarization model using an encoder-decoder recurrent neural network with attention. Specifically, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) cells were used in both the encoder and decoder layers. They built nine models by varying hyperparameters and reported Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores (Lin, 2004) to evaluate performance.

3 Methodology

3.1 Data Collection

A comprehensive dataset of Nepali news articles, was created with web scraping from various online news portals like BBC Nepali (Hasan et al., 2021) and others. A sample of the dataset obtained through this process is illustrated in Figure 1, and a link to the full dataset generated in this study is provided in the annexes.

3.2 Data Preprocessing

In this step, we have removed HTML tags, special characters, and irrelevant sections of the text (such as advertisements and navigation links). As the data was collected in two steps, the headlines and their corresponding article bodies had to be joined to create the complete dataset.

The collected dataset still had numerous characters that were not part of the Nepali Devanagari Character Set. These extraneous characters would have degraded the overall text quality and negatively impacted model performance. Specifically, the unwanted characters including Latin letters (a-z, A-Z), Arabic numerals (0-9), etc. To mitigate these issues, these characters have been removed from

the dataset. A prefix was also added to the start of each input text to indicate the summarization task to the model and it helped the model to better understand the context and the task it needs to perform.

learn the mapping from the input text to the target headlines during training. A data collator was also used after the tokenization, in order to dynamically pad the inputs and labels to the maximum length during the batching process, ensuring the efficient utilization of the model's capabilities.

S.N	Article	Headline
1.	काठमाडौँ — १३ औं सागमा नेपालले अहिलेसम्मकै ठूलो सफलता पाएपछि सरकारले खेलाडीलाई प्रोत्साहन गर्न पुरस्कार घोषणा गर्यो । पुरस्कार घोषणा गरेको ४० दिन बितिसक्यो । पदक विजेता खेलाडीहरुलाई सरकारले दिने भनेको पुरस्कार रकम हालसम्म पनि सरकारी प्रक्रियामै अल्झिरहेको छ । …	अर्थमन्त्रालय गएर रोकियो सागको पुरस्कार रकम
2.	काठमाडौं, पुस ६ गते । साताको दुई दिन लगातार रुपमा बढेको नेप्से बुधबार सामान्य अङ्कले गिरावट आएको छ । आज नेप्से परिसूचक ५.८९ अङ्क घटेर १८६७.२२ बिन्दुमा झरेको छ । त्यस्तै, सेन्सेटिभ इण्डेक्स १.३१ अङ्कले घटेको छ भने फ्लोट इण्डेक्स ०.११ अङ्क र सेन्सेटिभ फ्लोट इण्डेक्स ०.४० अङ्कले घटेको छ । …	नेप्से परिसूचक ५.८९ अङ्कले गिरावट
3.	काठमाडौँ — श्रीलंकामा हुने काभा पुरुष भलिबल च्यालेन्ज कपमा सहभागी हुने नेपाली खेलाडीलाई बिदाइ गरिएको छ । शुक्रबार काठमाडौँमा एक कार्यक्रमबीच खेलाडीहरुको बिदाइ गरिएको हो ।अर्को हप्तादेखि कोलम्बोमा सुरु हुने प्रतियोगितामा नेपालसहित श्रीलंका, माल्दिभ्स, बंगलादेश, अफगानिस्तान, साउदी अरब, तुर्केमिनिस्तान र उज्वेकिस्तान समावेश छन् । …	काभा भलिबल च्यालेन्ज कपमा सहभागी हुने नेपाली खेलाडीको बिदाइ
4.	रियो डे जेनेरियो — कोपा अमेरिकाको सेमिफाइनलमा दुई पुराना प्रतिद्वन्द्वी अर्जेन्टिना र ब्राजिल भिड्ने भएका छन् । अर्जेन्टिनाले शुक्रबारको क्वाटरफाइनलमा भेनेजुएलालाई २–० ले हरायो । त्यसमा लउतारो र जियोभानी लोले गोल गरे । आयोजक ब्राजिल यसअघि नै सेमिफाइनल पुगिसकेको छ । …	अर्जेन्टिना र ब्राजिल भिड्ने
5.	राष्ट्रपति विद्यादेवी भण्डारीले नेपालको संविधान २०७२ को धारा ७६ (२) अनुसार नयाँ सरकारको गठनका निम्ति राजनीतिक दलहरुलाई आह्वान गर्नुभएको छ। संविधानको धारा ७६ (१) अनुसार प्रतिनिधि सभामा कुनै पनि एक दलसित आवश्यक बहुमत नरहेका कारण राष्ट्रपतिले धारा ७६ को उपधारा (२) अनुसार नयाँ सरकार गठनको निम्ति आह्वान गर्नुको विकल्प थिएन। …	गठबन्धनको पक्षमा जनमत
6.	काठमाडौँ, भदौ २ गते ।बाढीपहिरोले क्षति पुर्याउने भन्दै मेलम्ची खानेपानी आयोजनाको पानी बन्द गरिएपछि काठमाडौँमा बागमतीको पानी वितरण गरिएको छ । बाढीपहिरोले विगतका वर्ष झैँ आयोजनाको हेडवक्स र सुरुङमा क्षति पुर्याउन नदिन पूर्वतयारीस्वरुप असारको पहिलो हप्तादेखि मेलम्चीको पानी वितरण प्रणाली बन्द गरिएको छ ।…	बागमतीको पानी वितरण
7	विश्वका हरेक भागको समग्र भूगोलमै विभिन्न प्राकृतिक प्रकोपको सम्भाव्य जोखिम छ । मानवीय क्रियाकलाप, प्राकृतिक सम्पदाको अविवेकी दोहन, जलवायु परिवर्तन आदिले प्राकृतिक प्रकोपको जोखिम बढाएको सर्वविदितै छ । प्राकृतिक प्रकोप न्यूनीकरण गर्न पूर्वतयारी तथा सतर्कता अपनाउन जरुरी छ ।…	विपत् न्यूनीकरणको पूर्वतयारी
8	बागलुङ — झन्डै तीन दशकअघि । जिल्लामा सडक सञ्जालमा समेत जोडिएको थिएन । भलिबल र फुटबललाई मात्रै खेल भनिन्थ्यो । मार्सल आर्टबारे जिल्लावासीमा थाहै थिएन । ललितपुरको ठेचोमा जन्मेका धनदास महर्जनले पहिलोपल्ट बागलुङ आएर मार्सल आर्ट्स चिनाए । २७ वर्षदेखिको उनको सक्रियताले अहिले बागलुङ मार्सल आर्ट्सको उत्कृष्ट जिल्लामा गनिन्छ। …	बागलुङमा जसले कराते सिकाए
9	जनकपुर — प्रदेश सरकारको अर्थ विविध शीर्षकमा रहेको ३ अर्ब ३४ करोड २३ लाख ८ हजार ३ सय ४१ रुपैयाँ आर्थिक वर्षको अन्त्यतिर विभिन्न मन्त्रालय र स्थानीय तहमा गरेको रकमान्तरमा 'चलखेल' भएको भन्दै विशेष संसदीय छानबिन समितिले मधेश प्रदेशका दुई मन्त्रीसँग आइतबार बयान लिएको छ ।…	३ अर्ब ३४ करोड रकमान्तर, मधेशका दुई मन्त्रीसँग संसदीय समितिको बयान
10.	काठमाडौँ — नेपाली कांग्रेसका महामन्त्री एवम् प्रतिनिधिसभा सदस्य गगनकुमार थापाले अहिले विश्वविद्यालयलाई राजनीतिक भागबण्डाको दलदलबाट निकाल्ने मौका रहेको बताएका छन् । त्रिभुवन विश्वविद्यालयका नवनियुक्त उपकुलपतिले रेक्टर र रजिष्टार नियुक्त गर्न प्रयास गरेकोमा प्रधामन्त्री पुष्पकमल दाहालले अवरोध गरेको सुनिएको भन्दै नेता थापाले त्रिविलाई राजनीतिक दलदलबाट बाहिर निकाल्ने उपकुलपतिको प्रयासमा अवरोध नबन्ने आग्रह गरेका छन् । …	प्रधानमन्त्रीलाई थापाको आग्रह- 'त्रिविलाई दलीय भाडबण्डाको दलदलबाट निकाल्न अवरोध नगर्नुहोस्

Figure 1: Data Sample

The input texts (articles) and the target texts (headlines) were then, tokenized to a maximum length of 1024 and 20 tokens respectively, ensuring that longer texts were truncated. The tokenized headlines from the previous step were then, set as labels in the model inputs. This helped the model to

3.3 Exploratory Data Analysis

The dataset, meticulously compiled from various news portals, encapsulated a total of 70,769 articles, categorized into ten distinct thematic areas: News, Sports, Opinion, Entertainment, Feature, Diaspora, World, Education, Blog and Others(Mix). The dataset had more data related to News cate-

gory, while blog category had the least amount of data. The average length of title and the text of the articles were found to be approximately 6 and 390 respectively. The dataset were, then splitted into training, validation, and test sets in an 70-20-10 ratio to ensure robust model evaluation.

Dataset type	Count
Training Set	49,538
Validation Set	14,154
Test Set	7,076

Table 1: Data distribution in training, validation and testing dataset

S.N	Category	Count
1	News	36798
2	Sports	18767
3	Others(Mix)	7258
4	Opinion	2358
5	Entertainment	2144
6	Feature	2014
7	Diaspora	750
8	World	462
9	Education	188
10	Blog	30
	Total	70,769

Table 2: Data Statistics

3.4 Model Selection and Fine-Tuning

3.4.1 Model Selection

In this study, we chose to use transformer-based models, specifically mBART (Liu et al., 2020) and mT5 (Xue et al., 2021), for abstractive summarization in Nepali. These models were selected over alternatives, due to their demonstrated effectiveness in multilingual settings and their ability to handle long-range dependencies in text as presented in (Taunk and Varma, 2023)(Baykara and Gungor, 2022)(Kahla et al., 2022). Both models have been pre-trained on large-scale multilingual datasets, making them particularly suitable for low-resource languages like Nepali, where language-specific data is limited.

- mBART: This model is a denoising autoencoder designed for multilingual machine translation and text generation. Its architecture is based on BART(Lewis et al., 2019), which reconstructs corrupted text sequences, allowing it to learn complex text representations across languages. We opted for mBART-large-50, which has around 600 million parameters, as it strikes a balance between performance and computational feasibility. Its ability to handle diverse languages makes it ideal for abstractive summarization in Nepali, where linguistic resources are scarce.

- mT5: As a text-to-text transformer model, mT5 is capable of handling a wide variety of NLP tasks, including summarization, translation, and classification. With 598 million parameters, mT5-base was selected due to its ability to perform multilingual tasks efficiently without requiring massive datasets for each language. The text-to-text approach allows for consistent handling of inputs and outputs, making it adaptable for low-resource languages like Nepali.

Both mBART and mT5 are well-suited for abstractive summarization because they generate new text rather than merely extracting parts of the source document, making them superior to earlier extractive methods. Given the size and complexity of these models, fine-tuning them with limited computational resources poses significant challenges. To address this, we incorporated two key techniques:

- Quantization (Dettmers et al., 2023): This technique reduces the precision of the weights in the model from 32-bit floating points to lower precisions, such as 4-bit or 8-bit. Quantization significantly reduces memory usage and accelerates computation by enabling faster arithmetic operations. In our study, 4-bit and 8-bit quantization was used for mBART, which allowed for a substantial reduction in computational cost without significantly compromising performance. This was crucial for making the model feasible to train in a low-resource setting.

- Low-Rank Adaptation (LoRA) (Hu et al., 2021): This method drastically reduces the number of trainable parameters by introducing low-rank updates to the model weights, rather than fully updating the entire model during fine-tuning. By applying LoRA, we were able to fine-tune large models like mBART and mT5 on Nepali text while using significantly fewer resources. This approach not

only made the fine-tuning process more efficient but also enabled faster convergence with fewer training steps.

Together, these techniques allowed us to fine-tune transformer models on relatively modest hardware, such as NVIDIA Tesla P100 GPUs provided by Kaggle, and enabled us to process our dataset efficiently for continuous 12hours.

3.4.2 Fine-Tuning

To enhance efficiency, we stored the dataset on Hugging Face. During the fine-tuning process, the model weights and configurations obtained after each training session were also pushed to Hugging Face for every model.

The following training arguments were set in the trainer and in the LoRA for the training in each models:

Parameters	Value
evaluation_strategy	epoch
learning_rate	5e-4
per_device_train_batch_size	5
per_device_eval_batch_size	5
weight_decay	0.01
num_train_epochs	3
per_device_train_batch_size	5

Table 3: Training arguments for trainer

Parameters	Value
r	32
lora-alpha	32
lora-dropout	0.1
bias	lora_only

Table 4: Training arguments for LoRA

The pre-trained models were then, adapted using the LoRA configuration. This involved updating the model's weights based on the low-rank adaptations, making it more efficient for the specific task of Nepali news headline generation. Finally, the adapted model were fine-tuned using the same training process as described earlier. The low-rank update enabled faster and more efficient training, resulting in a model that could generate high-quality headlines.

3.5 Evaluation

The evaluation strategy was set to run at the end of each epoch, allowing for periodic assessment of the model's performance during training. A custom function to compute evaluation metrics was provided to the trainer. This function calculated ROUGE scores to evaluate the quality of the generated headlines. The model's performance was finally assessed on the testing set using the custom evaluation function and helped in understanding the model's ability to generate accurate and coherent headlines from Nepali news articles.

To assess the models' performance, a survey was conducted with 62 participants, all of whom had at least 12 years of formal education in Nepali. They were asked to evaluate summaries of 10 different sentences from various categories for the evaluation. Each sentence had summaries generated by six different models. Participants were tasked with selecting the best summary overall based on criteria such as relevance, fluency, conciseness, informativeness, factual accuracy, and coverage.

4 Experimental Setup

For the execution of this experiment, the following setup was created:

4.1 Environment Configuration:

4.1.1 Hardware and Software Setup:

Given the substantial computational demands of fine-tuning our language models, we found Kaggle to be the most suitable platform. It offered free access to the NVIDIA TESLA P100 GPU (16GB), allowing us to conduct uninterrupted training sessions for up to 12 hours. For storing the data, the model weights and the configurations obtained after each training session, Hugging Face was used.

The experiments were ran using Python 3.12.3 along with key libraries such as PyTorch, BeautifulSoup, Selenium, Pandas, Numpy, Matplotlib, Plotly etc.

4.2 Experimental Workflow:

4.2.1 Dataset Handling:

The dataset was processed in batches of approximately 10,000 samples during training. For this experiment, 50,000 news articles along with their corresponding summaries were utilized for

training, while 14,000 were reserved for validation. At the start of each training session, the entire dataset was loaded into memory to facilitate efficient access for the models.

4.2.2 Batch processing and training time:

Batch processing was implemented to streamline training and evaluation. Training was performed with a batch size of 5 and ran for 3 epochs and validation was carried out at regular intervals to track performance improvements.

The total time taken to train each model was approximately 12 hours.

4.2.3 Optimization and Hyperparameters:

Hyperparameter tuning plays a vital role in optimizing the model's performance. While we set certain key hyperparameters such as learning rate (5e-4), weight decay (0.01), and batch size (5), additional tuning was performed to ensure optimal training efficiency.

- Learning Rate: The learning rate was selected based on experimentation. We observed that higher learning rates led to instability during fine-tuning, while lower learning rates slowed convergence. The value of 5e-4 was found to provide a good balance between fast convergence and model stability.

- Batch Size: A batch size of 5 was chosen due to memory constraints on the available GPUs. Larger batch sizes led to out-of-memory errors, while smaller batch sizes resulted in slower training. Using a batch size of 5 allowed for efficient utilization of GPU resources while maintaining training speed.

- Number of Epochs: We fine-tuned the model over three epochs, which was determined based on validation set performance. During early experimentation, we noticed that performance improvements plateaued after the third epoch, making additional epochs unnecessary.

4.3 Evaluation Setup:

4.3.1 Automated Evaluation:

In this study, we chose to use ROUGE (Recall-Oriented Understudy for Gisting Evaluation)(Lin, 2004) as the primary metric for evaluating the performance of the abstractive summarization models. Given the nature of our task—summarizing Nepali news articles—ROUGE is particularly well-suited for evaluating how well the models capture the key content of the source text. While additional metrics such as BLEU(Papineni et al., 2002) and METEOR(Lavie and Agarwal, 2007) could offer complementary insights, we determined that ROUGE alone provides sufficient coverage for the following reasons:

- Focus on Content Overlap: The goal of summarization is to ensure that the key ideas from the original text are preserved in the summary. ROUGE is highly effective in measuring this by quantifying the overlap of n-grams between the generated and reference summaries. This makes ROUGE particularly useful when the emphasis is on recall, as it ensures that the model does not miss critical information from the original text.

- Simplicity and Interpretability: ROUGE scores are widely accepted in the NLP community and offer a simple, interpretable way to measure performance. Introducing additional metrics may complicate the evaluation without necessarily providing new insights for the particular task of summarizing low-resource language texts like Nepali. The ROUGE metric's emphasis on recall and precision has proven reliable in many summarization tasks, and it correlates well with human judgment when the goal is content preservation.

- Alignment with Task Goals: The objective of this work is to generate coherent and concise summaries that faithfully represent the original content. Given that ROUGE scores provide a strong indicator of how much content overlap exists between the generated and reference summaries, they align well with our goals for content retention and accuracy. While BLEU and METEOR focus on fluency and sentence-level correctness, these aspects are already partially captured in human evaluation.

4.3.2 Human Evaluation:

For human evaluation, winner-takes-it-all approach was considered, where the human evaluators were asked to select the best summary overall based

on factors such as relevance, fluency, conciseness, informativeness, factual accuracy, and coverage among different summaries generated from different models for different sentences. A simple Google form was created and used to streamline the collection of feedback, ensuring that responses were gathered efficiently.

5 Results

The ROUGE scores for precision, recall, and F1-scores across all models are summarized in Table 6. These metrics provide a comprehensive evaluation of the summarization performance. Based on the results in the table, the 4-bit quantized mBART model with LoRA emerged as the best-performing model, consistently achieving the highest ROUGE scores in all categories. This indicates that the model was able to retain a higher degree of the original text's meaning while generating concise and fluent summaries.

Model	Number of votes received	Percentage of votes (%)
4bit quantized mBART + LoRA	235	34.06
8bit quantized mBART + LoRA	191	27.68
mBART + LoRA	164	23.77
mT5 + LoRA	100	14.49
4bit quantized mT5 + LoRA	000	00.00
8bit quantized mT5 + LoRA	000	00.00

Table 5: Results from the Human Evaluation

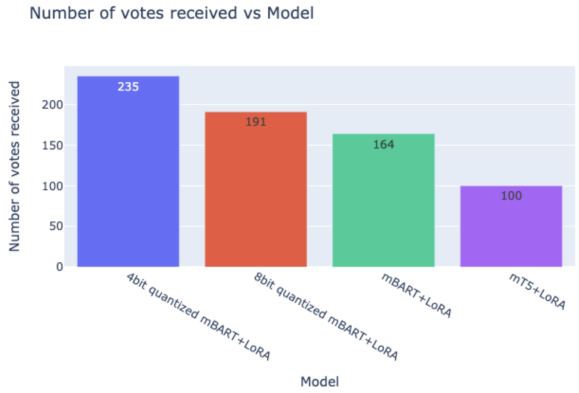

Figure 2: Results from the Human Evaluation

In addition to the automatic evaluation, the human evaluation results are presented in Table 5. These results further validate the performance of the 4-bit quantized mBART with LoRA, as it was selected the most (34.06%) by human evaluators. The model's summaries were consistently rated higher overall based on relevance, fluency, and factual accuracy compared to the other models. This strong alignment between the ROUGE scores and human preferences reinforces the model's ability to generate high-quality summaries that resonate with both automatic metrics and human judgment. The mT5 model, particularly in its quantized form, struggled with articles containing complex sentence structures or domain-specific terminology, such as those in the 'Politics' categories. This underperformance may be attributed to the model's sensitivity to quantization and its limited adaptation to the intricacies of the Nepali language. An example of a successful summary generated by mBART involved summarizing a political article where the model effectively captured the key points while maintaining fluency. In contrast, the mT5 model produced an incoherent summary with incomplete sentences, highlighting the model's challenges in handling complex topics.

mBART+LoRA	4bit quantized mBART+LoRA	8bit quantized mBART+LoRA	mT5+LoRA
पदक विजेता खेलाडीलाई दिने भनिएको पुरस्कार रकम अझै पनि सरकारी प्रक्रियामै अल्झिदै	पुरस्कार घोषणा गरेको ४० दिन बितिसक्यो, तर रकम अझै अल्झिरहेको छ : प्रशिक्षक न्यौपाने	पुरस्कार घोषणा ४० दिन ४० दिन बित्यो, अहिलेसम्म पनि रकमको छिनोफानो भएन	सरकारले दिने भनेको पुरस्कार रकम सरकारी प्रक्रियामै अल्झिरहेको
सामान्य अङ्कले घट्यो नेप्से	सामान्य अङ्कले घट्यो नेप्से	नेप्सेमा ५.८९ अङ्कको गिरावट	नेप्से परिसूचक ५.८९ अङ्कले गिरावट
काभा पुरुष भलिबल च्यालेन्ज कपमा सहभागी हुने नेपाली खेलाडीको बिदाइ	काभा पुरुष भलिबल च्यालेन्ज कपमा सहभागी हुने नेपाली खेलाडीको बिदाइ	काभा पुरुष भलिबल च्यालेन्ज कपमा सहभागी हुने नेपाली खेलाडीको बिदाइ	काभा पुरुष भलिबल च्यालेन्ज कप : नेपालताई बिदाइ
कोपा अमेरिका : अर्जेन्टिना र ब्राजिल भिड्ने	कोपा अमेरिका : सेमिफाइनलमा अर्जेन्टिना र ब्राजिल भिड्ने	कोपा अमेरिका : अर्जेन्टिना र ब्राजिल सेमिफाइनलमा	अर्जेन्टिना र ब्राजिल भिड्ने
राजनीतिक स्थायित्वको सुनिश्चितता	राजनीतिक स्थायित्वका चुनौती	राजनीतिक स्थायित्वको खाँचो	नयाँ सरकारको दाबी गर्ने दलहरु आह्वान

Figure 3: Summaries generated by different models (1-5)

Note: The highlighted entries in the above and below table received the maximum number of votes in the survey.

mBART+LoRA	4bit quantized mBART+LoRA	8bit quantized mBART+LoRA	mT5+LoRA
मेलम्चीको पानी वितरण सुरु	मेलम्चीको पानी काठमाडौंमा वितरण सुरु	बागमतीको पानी वितरण सुरु	मेलम्चीको पानी बन्द गरिएपछि काठमाडौंमा बागमतीको
प्राकृतिक प्रकोप न्यूनीकरणका उपाय	प्राकृतिक प्रकोपको जोखिम	प्राकृतिक प्रकोपको जोखिम न्यूनीकरण	विपत्को पूर्वतयारी
बागलुङ मार्सल आर्ट्सको उत्कृष्ट जिल्ला	' बागलुङ मार्सल आर्ट्सको उत्कृष्ट जिल्ला	'राजधानीमै बसेको भए यस्तो अवसर पाइने थिएन'	बागलुङमा करातेको चहलपहल
मधेशका दुई मन्त्रीसित छुट्टाछुट्टै बयान	मधेशका दुई मन्त्रीले लिए बयान	मधेश प्रदेशका दुई मन्त्रीसित छुट्टाछुट्टै बयान	मधेश प्रदेशका दुई मन्त्रीसँग विशेष संसदीय छानबिन समिति
विश्वविद्यालयलाई राजनीतिक भागबण्डाको दलदलबाट निकाल्ने मौका हो : महामन्त्री थापा	विश्वविद्यालयलाई राजनीतिक भागबण्डाको दलदलबाट निकाल्ने मौका छ : महामन्त्री थापा	अहिले विश्वविद्यालयलाई राजनीतिक भागबण्डाको दलदलबाट निकाल्ने मौका छ : महामन्त्री थापा	विश्वविद्यालयलाई दलीय भागबण्डाको दलदलबाट निकाल्ने मौ

Figure 4: Summaries generated by different models (6-10)

Model	ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F1	P	R	F1	P	R	F1
mBART+LoRA	0.3797	0.3517	0.355	0.211	0.196	0.1964	0.3684	0.3411	0.3443
4-bit quantized mBART+LoRA	0.3865	**0.354**	**0.359**	**0.2163**	**0.1984**	**0.1999**	**0.3754**	**0.344**	**0.3488**
8-bit quantized mBART+LoRA	**0.3871**	0.35	0.3574	0.2141	0.1941	0.1969	0.3754	0.3395	0.3466
mT5+LoRA	0.335	0.3248	0.3218	0.1746	0.1701	0.1675	0.3252	0.3154	0.3123

Table 6: ROUGE scores of different models on the test dataset

Note: The scores of the 4-bit quantized mT5+LoRA and 8-bit quantized mT5+LoRA models are not presented in the table as they produced zero scores across all calculated metrics, indicating that these configurations were not effective for the task at hand.

6 Conclusion

This study represents a significant step forward in addressing the challenges of abstractive summarization for low-resource languages like Nepali. By leveraging state-of-the-art multilingual transformer models, mBART and mT5, alongside innovative techniques such as Low-Rank Adaptation (LoRA) and quantization, the research successfully generated high-quality Nepali news headlines. The creation of a novel Nepali news dataset further supports the advancement of NLP resources for underrepresented languages.

The results demonstrated the superior performance of the 4-bit quantized mBART model with LoRA, which achieved high ROUGE scores and received the most favorable responses in human evaluations. This highlights its potential to deliver efficient and coherent summarization while addressing computational constraints. However, the mT5 model underperformed, indicating opportunities for further optimization tailored to Nepali's linguistic characteristics.

This work not only provides a practical framework for summarization in low-resource settings but also opens avenues for future exploration. Enhancements in quantization strategies, integration of diverse datasets, and the adoption of alternative evaluation metrics can further refine summarization models. Moreover, expanding this research to other South Asian languages can contribute to creating inclusive NLP tools that cater to diverse linguistic needs.

7 Limitations:

While this study provides significant insights into the potential of multilingual transformer models for abstractive summarization of low-resource languages like Nepali, it is not without its limitations. These constraints highlight areas where further improvements and investigations are necessary to enhance the effectiveness and applicability of the proposed methods. Below, we discuss the key limitations and outline directions for future research:

1. While LoRA and quantization techniques effectively reduce computational costs, their specific impact on linguistic characteristics, such as Nepali syntax and orthography, remains underexplored. Future studies could analyze how these techniques influence language-specific features and propose improvements for better adaptability.

2. The reliance on specific portals during dataset creation may have introduced domain bias, potentially limiting linguistic diversity. Expanding the dataset to include a wider range of sources across different domains could improve model generalization and adaptability in real-world applications.

3. The performance of mT5 models in this study underscores the need for customized fine-tuning and quantization approaches. Future research could experiment with advanced quantization levels, parameter-efficient tuning methods, or hybrid models tailored to Nepali's linguistic complexities.

4. While ROUGE scores were utilized effectively, additional metrics such as semantic coherence and logical consistency could enrich the evaluation. Future studies should employ more comprehensive metrics to better capture the quality and depth of model-generated summaries.

Acknowledgment

The authors thank the faculty members and staff of the Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, Tribhuvan University, Nepal, for their invaluable guidance and support.

References

Batuhan Baykara and Tunga Gungor. 2022. Turkish abstractive text summarization using pretrained sequence-to-sequence models. *Natural Language Engineering*, 29:1–30.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Preprint*, arXiv:1607.04606.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Preprint*, arXiv:1412.3555.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Mram Kahla, Attila Novák, and Zijian Yang. 2022. Fine-tuning and multilingual pre-training for abstractive summarization task for the arabic language. *Annales Mathematicae et Informaticae*, Accepted manuscript.

Rishi Khanal, Smita Adhikari, and Sharan Thapa. 2022. Extractive method for nepali text summarization using text ranking and lstm.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Kaushal Mishra, Jayshree Rathi, and Janardan Banjara. 2020. Encoder decoder based nepali news headline generation. *International Journal of Computer Applications*, 175:975–8887.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Robin Ranabhat, Amit Upreti, Bidhan Sangpang, and Shoaib Manandhar. 2019. Salient sentence extraction of nepali online health news texts.

Dhaval Taunk and Vasudeva Varma. 2023. Summarizing indian languages using multilingual transformers based models.

Bipin Timalsina, Nawaraj Paudel, and Tej Shahi. 2022. Attention based recurrent neural network for nepali text summarization. *Journal of Institute of Science and Technology*, 27:141–148.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. pages 606–615.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

A Appendix

A.1 Dataset Details

The dataset created as part of this study is available at the following link.
https://www.kaggle.com/datasets/dhakal2444/nepali-news-dataset

A.2 Models

- **Pre-trained models:**
  The pre-trained models used as part of this study is available at the following link.
  mBART: *https://huggingface.co/facebook/mbart-large-50*
  mT5:*https://huggingface.co/google/mt5-base*

- **Fine-tuned models:**
  The fine-tuned models created as part of this study is available at the following link.
  https://huggingface.co/collections/caspro/summarization-models-for-nepali-language-66c209bfac74db25dee47759