Kathmandu University
Journal of Science, Engineering and Technology

Multi-Stage Fine-Tuning of mT5-Small for Nepali News Summarization

Akarshan S. Shrestha, Albina Shrestha, Sophiya Shrestha, Yunika Bhochhibhoya, Arnav Bhatta, and Prakash Poudyal*

Department of Computer Science and Engineering, Kathmandu University, Nepal

Abstract

Nepali is a low resource language in the field of Natural Language Processing (NLP), with limited availability on labeled data and pre-training in language models. With the growing use of social media and large language models (LLMs) for concise news delivery instead of actual news platforms, there is a pressing need for methodologies for fine-tuning a usable and high-performing summarization model in low resource environments. In this work, we fine-tune the multilingual text-to-text transformer model, mT5-small, to perform abstractive, 2-3 sentence summarization of Nepali news articles. The model was trained on a custom dataset, which we collected extensively from Nepali news websites. Evaluation using ROUGE metrics yielded the scores ROUGE-1: 46.74, ROUGE-2: 33.01, ROUGE-L: 41.96 and evaluation using BERTScore-F1 gave the score 80.14, demonstrating viable performance for a low-resource language.

Keywords: News summarization; mT5; NLP; Nepali News; ROUGE; BERTScore.

1. Introduction

With the growing usage of social media, digital content has become the primary mode of news consumption for many people[1]. Increasingly, users are relying on platforms such as social media journalism pages and AI assistants like ChatGPT and Gemini to stay informed, rather than visiting traditional news websites [2]. This shift has created a need for news content that is concise, accessible, and engaging. It needs to be delivered in a format that aligns with the fast-paced habits of the modern generation.

This research was inspired by informal discussions with around 30 peers, who revealed that they predominantly consume news from social media pages like Routine of Nepal Banda and AawajNews or through AI chat assistants. Few reported visiting actual news websites, and when they did, it was infrequent. These conversations highlighted a clear gap that while news is abundant, its current presentation often fails to capture attention, and users rely on short and concise news sources. Hence, a solid and reliable text summarization approach can serve as a solution to this problem.

Text summarization in Natural Language Processing (NLP) offers a solution to this need. Broadly, summarization techniques fall into two categories: extractive and abstractive. Extractive methods identify and reuse important sentences or phrases from the source text, while abstractive methods generate new sentences while conserving the core context. Abstractive summarization is more human-like and challenging, especially for low-resource languages that lack large-scale datasets and pretrained models. This is because abstractive summarization requires the language model to learn how to create proper sentences with clarity and structure itself, while presenting the information required concisely. Despite these challenges, recent research has demonstrated that adapting pre-trained multilingual transformer models, such as mT5, for low-resource languages can yield highly competitive results in abstractive and extractive summarization tasks [3].

*Corresponding author. Email: prakash@ku.edu.np

While short abstractive summarization work such as headline generation has been done for the Nepali language, few works focus on producing 2–3 sentence summaries [4]. Addressing this gap requires not only building high-quality datasets but also identifying training strategies that maximize performance under resource constraints. The objective of this study is to develop and evaluate a multi-stage fine-tuning approach for the mT5-small model, tailored to Nepali news summarization, by leveraging a curated dataset pipeline that progresses from single-sentence to multi-sentence summaries. The study assesses the performance of the model using both lexical and semantic evaluation metrics and compares the results with existing work.

The rest of the paper is organized as follows. Section II reviews related work on mT5, multilingual summarization, and prior research in Nepali NLP, highlighting the major gaps our study addresses. Section III outlines the methodology, including dataset collection, model choice, and the multi-stage fine-tuning approach. Section IV presents the experimental results and compares the performance of the model with existing studies. Section V concludes the paper with a discussion on practical applications and directions for future research.

2. Literature Review

Transformer-based architectures have revolutionized NLP, enabling state-of-the-art performance in various tasks, including summarization. Among them, Google's mT5 model extends the T5 architecture to 101 languages, trained on the mC4 corpus [5]. mT5 uses a text-to-text framework where all tasks, including summarization, are framed as sequence-to-sequence generation. Its multilingual nature and transfer learning capabilities make it highly suitable for low-resource languages such as Nepali. However, a key gap that remains is the development of optimal and resource-efficient fine-tuning strategies for complex, language-specific tasks such as generating coherent, multi-sentence, abstrac-

tive summaries in low-resource languages [3].

With the rise of transformer models and LLMs, several research works have been conducted in the field of NLP for Nepali language as well as other low resource languages in recent times. [6] presents XL-Sum, a large-scale multilingual abstractive summarization dataset covering 44 languages, including the Nepali language and several other languages, thus enabling and facilitating research on summarization in traditionally under-served languages. [7] presents significant works in Nepali NLP by creating a large text corpus and pre-trains three transformer models (BERT, RoBERTa, and GPT-2) on it. It achieved state-of-the-art results on Nepali language understanding and text generation tasks. [4] works on training an abstractive model using quantized LoRA training configurations for multilingual models mT5 and mBART. Their model was trained to generate headlines using a large collection of scraped Nepali articles dataset from the web. Similarly, [8] introduced a NepBERTa, a transformer-based model, which achieved better performance on several NLP tasks for Nepali language, such as Named Entity Recognition (NER), Content Classification, parts-of-speech (POS) tagging, and Categorical Pair Similarity.

Outside the Nepali context, mT5 has been successfully fine-tuned for abstractive summarization in other low-resource languages [3] [6] [9]. Nasution et al. [10] and Singh & Choudhary [9] proposed multi-stage or hybrid fine-tuning strategies. These works show that hybrid or multi-stage approaches for training an abstractive model can give better results compared to linear approaches, showing improvements in factual accuracy and fluency. However, these strategies have not been systematically explored for Nepali, nor have they been optimized for multi-sentence abstractive summarization. Our work addresses this gap by applying a hybrid, multi-stage training strategy inspired by these prior studies.

3. Methodology

3.1. Dataset Collection

For constructing the dataset, we implemented custom scraping scripts using Python and BeautifulSoup. We targeted popular Nepali news websites such as Setopati, OnlineKhabar, and Ekantipur. Each scraper was tailored to handle the site-specific layouts and content blocks. Over 5000 full-length articles were scraped and saved from each website, including metadata such as publication date, source, and raw text.

The preprocessing pipeline handled removal of HTML artifacts, redundant headers, and formatting noise, as well as genre classification for organizing content into categories. We collected datasets in two forms – a lead sentence dataset and a highlights dataset.

Firstly, we identified that news articles have always had a summarized structure in their first sentences – a standard practice in journalism. We extracted the first sentence of about 5000 articles to create the first-sentence dataset. However, to to prevent the model from learning to copy the first sentence, we removed the original first sentence or introductory paragraph from the article input.

Secondly, we extracted highlight sections from some articles, which served as reference summaries for training. The goal was to create a model that creates a 2-3 sentence summary of a news article that is brief, clean and captures the key details of the article. While scraping we noticed that certain articles from OnlineKhabar and Ekantipur contained a highlights section containing key details in the article. The highlights were AI-generated and reviewed by humans for OnlineKhabar, and human-written for Ekantipur. We extracted these articles and their highlights section as the high quality dataset.

3.2. Model Choice

We chose the mT5-small model from Google [5], designed for multilingual sequence-to-sequence tasks. Given the constraints of limited computational resources, mT5-small which contains approximately 300M parameters offered a good balance of performance and feasibility. It is already trained on a massive multilingual corpus, the mC4, and given the multilingual nature of mT5 and its prior training on various language tasks, it was a suitable candidate for adapting to the Nepali language with relatively limited data.

3.3. Multi-Stage Summarization Strategy

The training pipeline was divided into three distinct stages:

Stage 1: Lead Sentence Dataset – Since the standard convention by journalism practices on creating highly relevant first sentences in news articles, we found it to start training on the lead sentence dataset before moving on to the datasets with 2–3 sentences, even though the goal was to create a dataset intended to generate 2–3 sentences. We discovered that the model develops better fluency and sentence structure when first trained on this dataset before moving on rather than directly being trained on the 2–3 sentence datasets. The model also learns the importance of the context given by the lead sentence for further information density, a trait that would be crucial in the future stages for a good 2–3 sentence output.

Stage 2: Highlights Dataset – After that, fine-tuning on high-quality highlight-summary pairs was done to increase information density along with factual accuracy. It is important to note that training on this stage must be longer, i.e., for more epochs than for stage 1, otherwise the model develops a bias to develop a single sentence, which is expected behavior.

Stage 3 – With the strong foundation laid out by stages 1 and 2, further training on a portion of the public dataset 'Someman/news_nepali'[11] from Hugging Face was done to generalize the model to diverse writing styles. We used this dataset due to time constraints which limited the size of the other datasets. Also, the highlights dataset for Ekantipur did not include sentences but rather points, so it was important to train the model on actual sentence summaries as well.

3.4. Training Setup and Configuration

We used the Hugging Face Hub for model checkpointing, with Weights & Biases (wandb.ai) for experiment tracking, loss monitoring, and performance visualization.

The training conditions for each stage are as follows:

Stage 1 and 2: Learning Rate: 2e-05, Optimizer: AdamW, Training environment: Local Machine with GTX1660ti. To further optimize training efficiency, we employed Low-Rank Adaptation (LoRA) during Stage 1 and Stage 2. LoRA reduces the number of trainable parameters by injecting low-rank matrices into the attention layers of the model. This enabled us to fine-tune the model with reduced memory consumption and faster convergence, making it ideal for training on limited hardware of the local machine.

Stage 3: Learning Rate: 5e-06, Optimizer: AdamW, Training environment: Kaggle Notebooks with GPU T4 x2. LoRA was not used in Stage 3, as full fine-tuning was feasible due to the T4×2 GPU setup provided by Kaggle.

4. Results and Discussions

4.1. Evaluation Metrics

Table 1 assesses the performance of the fine-tuned abstractive summarization model for Nepali news articles for different stages.

Table 1: ROUGE and BERTScore Results from Different Training Stages

Metric	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore-F1
Base mT5-small	19.89	7.51	15.74	15.77	61.11
Stage 1+2	22.27	6.89	13.62	13.56	73.09
Stage 3	**46.74**	**33.01**	**41.96**	**41.95**	**80.14**

Table 2: Comparison of Summarization Models

Metric	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F1
Dhakal & Baral	35.90	19.99	34.88	NA
XL-Sum	28.70	17.26	26.23	75.68
Our model	**46.74**	**33.01**	**41.96**	**80.14**

Table 3: Example Summaries Generated by Our Model

Article (excerpt)	Generated Summary
२३ असार, काठमाडौं । नेपाल शिक्षक महासंघविरुद्ध संसदीय समितिमा उजुरी परेको छ । शिक्षक नारायणप्रसाद घिमिरे र कार्यालय सहयोगी शंकरबहादुर कुमालले महासंघ विरुद्ध प्रतिनिधि सभाको शिक्षा, स्वास्थ्य तथा सूचना प्रविधि समितिमा उजुरी दिएका हुन् । उजुरीमा महासंघ दर्ताविना नै सन्चालन भइरहेको भनिएको छ । दर्ता विना संस्था सञ्चालन गरी लेटरप्याड तथा छापको समेत अनाधिकृत, प्रयोग गर्दै सहयोग रकम, सदस्यता शुल्क, नवीकरण दस्तुरका नाममा रकम उठाएको आरोप उजुरीकर्ताले लगाएका छन् । महासंघको गतिविधि रोकेर आवश्यक अनुसन्धान गर्न निवेदककर्ताले माग गरेका छन् ।	शिक्षक महासंघविरुद्ध संसदीय समितिमा उजुरी परेको छ । शिक्षक नारायणप्रसाद घिमिरे र कार्यालय सहयोगी शंकरबहादुर कुमालले महासंघको गतिविधि रोकेर आवश्यक अनुसन्धान गर्न निवेदककर्ताले माग गरेका छन् ।
लन्डन — आर्सनलले आइतबार स्पेनिस मिडफिल्डर मार्टिन जुबिमेन्डीलाई रियल सोसिडाडबाट अनुबन्धित गरेको जनाएको छ । उनको सरुवा रकम ७ करोड ५० लाख डलर भएको बताइएको छ । स्पेनबाट युरो २०२४ को उपाधि जितेका जुबिमेन्डीलाई यसअघि लिभरपुल र रियल म्याड्रिडले पनि लिन तत्परता देखाइरहेका थिए । 'यो मेरो खेल जीवनका लागि निकै ठूलो क्षण हो,' डिफेन्सिभ मिडफिल्डर जुबिमेन्डीले भने । उनले आर्सनलसँग पाँचवर्षे सम्झौतामा हस्ताक्षर गरेको जनाइएको छ । रियल सोसिडाडले उनका लागि तोकेको रकमभन्दा बढीमा आर्सनलले जुबिमेन्डीलाई लिएको एथ्लेटिकले जनाएको छ । जसअनुसार विभिन्न किस्तामा उनको सरुवा रकम आर्सनलले बुझाउनेछ । 'यो यस्तो क्षण हो, जुन मैले प्रतीक्षा गरिरहेको थिएँ, जसलाई पूरा गर्न चाहन्थें । यो कति ठूलो क्लब हो भन्ने यहाँ खुट्टा टेकेपछि मात्र थाहा हुन्छ । आर्सनल नै मैले रोज्नुको कारणचाहिँ उनीहरू जसरी खेल्छन्, त्यो मेरो खेललाई सुहाउँदो छ । उनीहरूले पनि हालै आफ्नो क्षमता देखाइसकेका छन् । राम्रो नतिजा अझै आउन बाँकी छ ।' २६ वर्षे जुबिमेन्डी सोसिडाडको एकेडेमीबाट अगाडि आएका खेलाडी हुन् र उनले यो क्लबका लागि २ सय ३६ म्याच खेलिसकेका छन् । उनले स्पेनको राष्ट्रिय टोलीबाट पनि १९ पटक खेलिसकेका छन् । थोमस पार्टेय र जोर्गिन्हो क्लब छाडेर गएकाले आर्सनलको मिडफिल्डमा जुबिमेन्डीले ठूलो प्रभाव पार्ने विश्वास व्यवस्थापक मिकेल आर्टेटाको छ । जुबिमेन्डीको आगमनपछि आर्सनलका कीर्तिमानी अनुबन्धित डेक्लन राइस र कप्तान मार्टिन ओडेगार्डले अझ आक्रामक भएर खेल्न पाउने छन् । आउँदो सिजनका लागि आर्सनल भित्रिने उनी दोस्रो खेलाडी हुन् । टोलीले गोलरक्षक केपा एरिजाबालागालाई चेल्सीबाट भित्र्याइसकेको छ ।	आर्सनलले स्पेनिस मिडफिल्डर मार्टिन जुबिमेन्डीलाई रियल सोसिडाडबाट अनुबन्धित गरेको जनाएको छ । उनको सरुवा रकम ७ करोड ५० लाख डलर भएको थियो ।
२३ असार, काठमाडौं । पूर्वराजा ज्ञानेन्द्र शाहले आज आफ्नो ८० औं जन्मदिन मनाएका छन् । निर्मल निवासमा आयोजित एक विशेष कार्यक्रममा ज्ञानेन्द्रले आफ्नो जन्मदिन मनाएका हुन् । उनको जन्मदिनमा सहभागी हुन पुग्ने समर्थकहरूको बाक्लो भिड थियो । समर्थकहरूले लाइन बसेरै पूर्वराजा शाहलाई शुभकामना दिए । जन्मोत्सवको अवसरमा शाहले अहिले नेपाल र नेपालीका लागि प्रार्थना गर्ने समय भएको बताए । पूर्वराजा शाहले नेपालमा जन्मिएर मात्र नेपाली नहुने भन्दै नेपाल नै आफूभित्र समाहित हुनुपर्ने र त्यसपछि मात्रै नेपालको माया जाग्ने पनि बताए ।	ले आज आफ्नो ८० औं जन्मदिन मनाएका छन् । उनको जन्मदिनमा सहभागी हुन पुग्ने समर्थकहरूले लाइन बसेरै पूर्वराजा शाहलाई शुभकामना दिए ।

The primary goals were to evaluate the accuracy, fluency, and semantic similarity of the generated summaries compared to reference summaries. For that, we conducted both ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BERTScore evaluations. ROUGE measures lexical overlap between generated and reference summaries. BERTScore evaluates semantic similarity using contextual embeddings from the BERT model.

We constantly experimented with different checkpoints to continue from after each stage of the training. For example, stage 1 training loss flattened after 5 epochs, hence giving 5 checkpoints to continue stage 2 training from. However, taking the best trained 5-epoch checkpoint from stage 1 would result in an early overfitting from stage 2 onwards, with the resultant model heavily biased to create only single sentence summaries, and that too with lower evaluation scores in the end.

Our best result was obtained by training first in stage 1 by 2 epochs, proceeding with stage 2 by 3 epochs and finally training in stage 3 by 3 epochs. The model can be retrieved from https://huggingface.co/aku47z/mt5-small-nepali-v3.1.

The results from Table 1 show a significant improvement in both contextual overlap and semantic similarity in the final-stage model. The improvement in BERTScore by the end of stage 2 reflects the model's ability to grasp important contexts. It is important to note that the reduced ROUGE scores do not reflect a reduc-

tion in the quality of summaries after stage 1 and 2. This is due to the highly abstractive nature of the model resulting from training on lead sentences from articles without first sentences or paragraphs, as well as highlights. Also, the summaries produced by the base mT5-small often lacked some if not all characteristics of sentence structure and clarity, which was fixed by the end of stage 1 and 2. The significant improvements in scores in stage 3 indicate a strong foundation from stage 1 and 2. The improved ROUGE scores in particular indicate the model's improved ability to capture multi-word sequences from the articles, which it learned from the diversified dataset in stage 3.

Based on our experiments, we also believe that an excellent one sentence summarizer model can be trained from the lead sentence dataset if trained for longer and on an expanded dataset.

4.2. Comparative Performance Analysis

Table 2 presents a comparison of our best-performing model against two past works on Nepali abstractive summarization. The results for the XL-Sum model [6] were re-evaluated on our dataset for a direct comparison, while the scores for the Dhakal & Baral model [4] are taken directly from their paper. As the original Dhakal & Baral paper did not report a BERTScore, this metric is omitted for their model to reflect the original findings.

The results presented in Table 2 highlight the performance of our multi-stage fine-tuning approach compared to other studies. While a direct comparison with previous studies is challenging due to variations in datasets and task definitions, our model's performance on multi-sentence abstractive summarization is noteworthy. Specifically, our mT5-small model, trained with the proposed multi-stage strategy, achieved superior ROUGE and BERTScore-F1 scores compared to the XL-Sum Baseline, which utilized a larger mT5-base model. This finding is significant for low-resource NLP, as it demonstrates that a well-designed fine-tuning methodology can enable a smaller model to perform well, and validates the resource-conscious approach of this study. Approaches like this can be invaluable for low-resource environments.

4.3. Qualitative Analysis

Some examples of the summaries generated by our multi-stage fine-tuned model are shown in Table 3. Manual inspection of the summaries show that the model was able to condense article content into fluent, grammatically correct, 2–3 sentenced summaries, all while preserving factual information from the original articles.

However, in some edge cases (last example in Table 3), the model occasionally omitted specific names or subjects from the sentence. This limitation is consistent with other summarization systems, especially in low-resource languages. These can be overcome with more consistent, higher quality and larger datasets for training and stricter human evaluation.

5. Conclusion

This paper presented the design, development and evaluation of a Nepali language summarization model by leveraging a fine-tuned mT5-small transformer model. Our system demonstrated the feasibility of performing abstractive summarization in a low-resource language setting, yielding strong performance on ROUGE and BERTScore metrics. Through our implementation of a multi-stage training pipeline and with strategic training and implementation under low resources, it is possible to build language training methodologies that can fine-tune pre-trained models to perform effectively in low resource scenarios.

Beyond its academic contribution, this work has significant practical applications. Summarization models like these can be further enhanced and used to automatically generate concise summaries for news aggregation platforms and to improve information retrieval for Nepali search queries. Furthermore, this model can enhance accessibility to information for native Nepali speakers and support content generation for social media platforms.

Future work can focus on: incorporating reinforcement learning with human feedback (RLHF) for improved multi-stage training strategies, fine-tuning larger models like mT5-base for comparison, expanding and training with more diverse genres and parallel summaries in the dataset, and training a versatile summarization model that can perform abstractive summarizations of varying lengths and tones.

Acknowledgment

The authors would like to thank the contributors to the Nepali news websites which were accessed for data sources in this research. The authors would also like to thank the reviewers who provided constructive feedback and gave their valuable time to make this paper possible.

References

[1] Shakya N & Poudyal P, Detection of fake news using deep neural networks, *Journal of KUSet*, 16(2). URL https://nepjol.info/index.php/KUSET/article/view/62625.

[2] Newman N & Cherubini F. Journalism, media, and technology trends and predictions 2025. Tech. rep., Reuters Institute for the Study of Journalism (2025).

[3] Munaf M, Afzal H, Iltaf N & Mahmood K. Low resource summarization using pre-trained language models (2023). https://doi.org/10.48550/arXiv.2310.02790. URL https://doi.org/10.48550/arXiv.2310.02790.

[4] Dhakal P & Baral D S. Abstractive summarization of low resourced nepali language using multilingual transformers (2024). https://doi.org/10.48550/arXiv.2409.19566. URL https://doi.org/10.48550/arXiv.2409.19566.

[5] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Saladi A, Gehring S, Herzig J & Stahlberg F, mt5: A massively multilingual pre-trained text-to-text transformer, *arXiv preprint arXiv:2010.11934*. URL https://arxiv.org/pdf/2010.11934.

[6] Hasan T, Bhattacharjee A, Islam M S, Samin K, Li Y F, Kang Y B, Rahman M S & Shahriyar R. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages (2021). URL https://arxiv.org/abs/2106.13822.

[7] Thapa P, Nyachhyon J, Sharma M & Bal B K. Development of pre-trained transformer-based models for the nepali language. In: *Proceedings of the International Committee on Computational Linguistics* (2025), pp. 9–19. URL https://aclanthology.org/2025.chipsal-1.2.pdf.

[8] Timilsina S, Kafle G & Joshi H. Nepnerta: Nepali named entity recognition transformer architecture. University of Aberdeen Repository (2024). URL https://aura.abdn.ac.uk/handle/2164/21465.

[9] Singh M & Choudhary S, An improved two-stage approach for abstractive text summarization, *International Journal of Emerging Trends in Engineering Research*, 8(5) (2020) 2529–2535. URL http://www.warse.org/IJETER/static/pdf/file/ijeter1058102020.pdf.

[10] Nasution S, Ferdiana R & Hartanto R, Towards two-step fine-tuned abstractive summarization for low-resource language using transformer t5, *International Journal of Advanced Computer Science and Applications*, 16(2).

[11] Someman. news_nepali: A Nepali News Summarization Dataset. https://huggingface.co/datasets/Someman/news_nepali (n.d.).