Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online) 2026, Vol. 5, No. 1

Nepali News Headline Generation using mBART Model

Bibek Prasad Paneru
National College of Engineering, Tribhuvan University, Nepal.

Mohit Budhathoki
National College of Engineering, Tribhuvan University, Nepal.

Sumit Panta
National College of Engineering, Tribhuvan University, Nepal.

Arudhi Bohora
National College of Engineering, Tribhuvan University, Nepal.

Divya Bhattarai
National College of Engineering, Tribhuvan University, Nepal.

Sharmila Bista*
National College of Engineering, Tribhuvan University, Nepal.
sharmila@nce.edu.np

*Correspondint author

***Abstract*—With the rapid increase in digital news consumption, generating concise and informative headlines has become essential in the present world. Social media users are increasingly getting news from their respective platforms. This study uses mBART, a multilingual transformer-based model, to automatically generate headlines for Nepali news articles. For effective fine-tuning, the model was trained on 86,628 news articles scraped from various Nepali news portals, utilizing a sequence-to-sequence architecture and LoRA. By using a self-attention mechanism, the model captures more context and performs better than conventional methods. To guarantee data quality, it was first subjected to filtering, pre-processing, and tokenization. The model performed well in capturing the structure and relevance of the content, as evidenced by its 0.4545 ROUGE scores. Flutter was used to create an intuitive user interface that made it possible to input and view generated headlines with ease. Social networks, content aggregation platforms, and news portals can all incorporate this research. The work promotes automation in digital journalism and advances Nepali natural language processing.**

***Keywords*—Digital Journalism, Fine-tuning, Flutter, LoRA, mBART, ROUGE, Self-attention Mechanism, Social Media, Web-scraping**

I. INTRODUCTION

News Headline Generation refers to the process of creating a brief and engaging title that encapsulates the core message of a news article. This procedure is essential since headlines provide readers with a quick overview of an article's content and serve as their first point of engagement. The goal of Nepali News Headline Generation using the mBART Model is to distill long texts into summaries while maintaining the essential information and meaning. Large amounts of news content are processed effectively by this technology in Nepali, resulting in headlines that are closely matched summaries. To ensure a broad range of topics and styles are covered, the system uses web-scraping techniques to gather large datasets from various Nepali news sources.

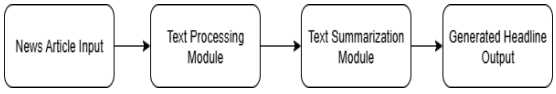

Figure 1: System Block Diagram

Traditional summarization techniques, such as term frequency, TF-IDF, and graph-based algorithms like TexRank and LexRank, played a foundational role by leveraging shallow linguistic features and statistical heuristics to identify important content within a document [1]. These methods frequently fall short in capturing the nuance of human language, leading to summaries that could be inconsistent and incoherent. Conversely, the mBART model emphasizes how deep learning transformed Natural Language Processing (NLP). It is especially useful for complicated languages like Nepali because of its pre-training on big datasets, which allows it to pick up nuances and patterns unique to the language.

In Natural Language Processing (NLP), text summarizing is an essential process that seeks to preserve the essential information while distilling vast amounts of text into concise, relevant summaries. Effective summarizing methods, both extractive and abstractive, are now crucial for efficient information retrieval and comprehension due to the explosive growth of digital content [2]. The traditional methods established the theoretical and algorithmic groundwork for modern approaches by demonstrating how surface-level features and unsupervised techniques can be effectively used for summary generation. However, the shortcomings of extractive methods, such as repetition, a lack of abstraction, and incoherent summaries, led the way for a shift toward neural and deep learning-based models, which try to emulate human-like comprehension and summarization [3].

II. RELATED WORK

Text summarization has evolved significantly over the years, with early works primarily focused on extractive methods. From the original news articles, the seq2seq LSTM model successfully produced succinct and pertinent news headlines. The generated headlines maintained good quality in terms of fluency, informativeness, and resemblance to human-written headlines, according to the ROUGE and BLEU scores. When it came to capturing the primary idea of news items, abstractive summarization outperformed extractive strategies for creating headlines. The work demonstrates that deep learning-based sequence models can be effectively used for automatic headline generation in Nepali and other low-resource languages [4]. Babar et al.(2013) present an overview of text summarization approaches, highlighting key concepts, types, and challenges of properly summarizing big textual data. Their findings highlighted the benefits

DOI : https://doi.org/10.3126/jost.v5i1.93035

Article History: Received: 1 Nov. 2025, Revised: 8 Jan. 2026, Accepted: 17 Feb. 2026.

and drawbacks of extractive summarization, which selects key sentences directly from the source text [5]. Based on these ideas, Khanal et al. (2022) suggested an extractive summarization strategy for Nepali literature that merged standard text ranking algorithms like TextRank with Long Short-Term Memory (LSTM) networks. When tested on a Nepali news corpus, their model received a ROUGE-1 F1 score of around 0.45, indicating a 10-15% improvement over traditional statistical methods by applying deep learning for greater contextual awareness. [6]. Similarly, Patil et al. (2022) created a general-purpose extractive summarizer with NLP techniques like part-of-speech tagging and named entity recognition. Their system achieved a ROUGE-2 recall score of around 0.40 across multiple domains, indicating effective sentence selection but limited abstractive ability due to its extractive nature [7].

With the emergence of deep learning, the focus switched to abstract methods. Mishra et al. (2020) developed an encoder-decoder architecture with attention mechanisms specifically designed for Nepali news headline production. Their model earned a BLEU score of roughly 0.30, indicating that it can generate coherent headlines despite morphological and syntactic difficulties, though further improvements would require bigger annotated datasets [8]. Similarly, multilingual transformer models marked a significant step forward. Dhakal and Baral (2024) used pre-trained multilingual transformers (mBART and mT5) for Nepali abstractive summarization. Their model achieved ROUGE-1 F1 and ROUGE-L scores of around 0.52 and 0.48, respectively, exceeding earlier LSTM and extractive models by effectively capturing semantic and long-range dependencies [9]. The field of NLP for low-resource languages has benefited greatly from the creation of the Nepali news headline summarization corpus and the improvement of multilingual models. Bhuktiyarov and Gusev (2020) improved transformer-based models, including BART and T5, for real-time headline generation. Their approach outperformed typical sequence-to-sequence models in creating fluent, contextually relevant outputs on Nepali headlines, with a BLEU score of around 0.35 [10]. Li et al. (2021) developed the HG-News model, which uses generative pre-training and domain-specific fine-tuning to increase headline coherence and relevance. On multilingual datasets, including Nepali, HG-News achieved a ROUGE-2 F1 score of 0.44 and a METEOR score of around 0.33, indicating significant gains over baseline transformer models [11].

Liu et al. (2022) introduced the CPT framework, which allows mBART to continuously learn from multiple datasets. This strategy enhanced adaptability to low-resource languages such as Nepali, resulting in ROUGE-1 and BLEU scores of around 0.50 and 0.32 in headline generation tasks, despite little Nepali data. [12]. Timalsina, Paudel, and Shahi(2022) suggest using an attention-based recurrent neural network (RNN). A seq2seq architecture with an attention mechanism allows the model to concentrate on the most pertinent portions of the input text, resulting in summaries that are both logical and educational. In order to train and assess the model, the authors created a Nepali news corpus. They showed that, in contrast to conventional RNN-based seq2seq models, attention greatly enhances summary quality with Rouge Score 0.44 [13]. Panthaplackel et al. (2022) created a system that uses extractive and abstractive methods to generate headlines in updated articles. The Headline Refinement Evaluation Network (HREN) benchmark revealed that their model earned a ROUGE-2 F1 score of around 0.42, efficiently balancing content relevance and linguistic quality in dynamic news contexts. [14]. Awasthi et al. (2021) present a comprehensive survey on NLP-based text summarization, categorizing methods into abstractive and extractive approaches. It was found that extractive techniques like TF-IDF, TextRank, and Word2Vec remain useful for keyword extraction and sentence rating. While extractive approaches typically score between 0.40 and 0.48 on ROUGE-1, abstractive and hybrid models outperform them by producing more cohesive and human-like summaries. The authors draw the conclusion that although deep learning and transformer designs have advanced significantly, multilingual summarisation, domain-specific applications, and enhanced assessment methods should be the key areas of future study [15].

III. METHODOLOGY

The methodology adopted in this study is designed to ensure a robust and reproducible framework for the development of the system. It involves collecting textual data through web scraping, followed by preprocessing to enhance data quality. The processed dataset is divided into training, validation, and test sets for model development and evaluation. The model is trained, fine-tuned, and tested to ensure optimal performance, and finally deployed through an interface for practical application.

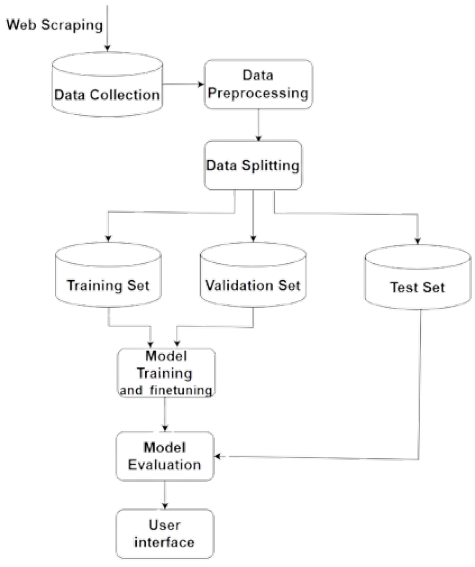

Figure 2: System Workflow Diagram

A. Data Collection

Data was scraped from the Nepali news portal Online Khabar, BBC Nepali, and Ratopati, resulting in a dataset of 86,628 rows and 3 columns. For web scraping, *libraries* like BeautifulSoup were used. Data in each portal was in a different format, and different strategies had to be adopted to extract data from them. The data were filtered to remove duplication and stored in CSV format containing the article text and a human-written headline. A sample of the dataset obtained from this process has been presented in Table I.

Table I: Sample Dataset with Nepali News Headlines

URL	News Article	Headline
https://www.ratopati.com/story/426873/hospital	काठमाडौं । शिक्षण अस्पताल महाराजगन्जले करार सेवामा पदपूर्ति गर्न विभिन्न पदमा आवेदन आह्वान गरेको छ । अस्पतालका लागि चिकित्सा शास्त्र अध्ययन संस्थान कार्यान्वयन समितिको स्वीकृत अस्थायी दरबन्दीमा क्लिनिकल रजिष्ट्रार पदका लागि करार सेवामा पदपूर्ति गर्न लागिएको सूचनामा उल्लेख छ ।	शिक्षण अस्पताल महाराजगन्जले मागयो विभिन्न पदमा आवेदन
https://www.ratopati.com/story/229112/school,	चिकित्सा शिक्षा आयोगले स्नातक तहका विभिन्न शैक्षिक कार्यक्रमहरूका लागि एकीकृत प्रवेश परीक्षा काठमाडौंमा संचालन गर्ने बारे पुनः जानकारी गराएको छ । एकीकृत प्रवेश परीक्षाको परीक्षा केन्द्रसम्बन्धी कही उम्मेदवारबाट दुविधा भएको भन्दै आयोगले सूचनामार्फत चैत्र १८ गतेबाट काठमाडौंमा सञ्चालन गरिने जानकारी गराएको हो ।	चिकित्सा शिक्षाका प्रवेश परीक्षा काठमाडौंमा हुने
https://rb.gy/0qujhe	काठमाडौं । त्रिभुवन विश्वविद्यालयको नीति अधिवेशनबाट ११ बुँदे प्रस्ताव पास भएको छ । सभामा पेश भएका १२ बुँदे प्रस्तावमध्ये ११ बुँदा पास भएका हुन्। प्रधानमन्त्री केपी शर्मा ओलीले पुस ७ गते बिहान ११ बजे नीतिबारे थप छलफलका लागि बैठक बोलाएका छन् ।	त्रिभुवन विश्वविद्यालयको नीति अधिवेशनबाट ११ बुँदे प्रस्ताव पारित

B. Data Preprocessing

The scraped data was cleaned, removing HTML tags, special characters, and symbols. In particular, Arabic numbers (0–9) and Latin letters (a–z, A–Z) are among the undesirable characters. These characters have been eliminated from the dataset in order to lessen these problems. In the tokenization process, the source article is truncated to a maximum length of 512 tokens, while the target headline is limited to 50 tokens.

C. Model Training and Fine-Tuning

We selected the multilingual transformer models mBART-large-50 with around 600M parameters for our Nepali summarization because they provide robust multilingual comprehension, flexibility in low-resource environments, and controllable processing demands in contrast to larger versions. Since it was still expensive to fully train these models, we combined LoRA with QLoRA's quantization approaches to lower resource consumption and trainable parameters without sacrificing performance [16].QLoRA enables memory-efficient fine-tuning of a quantized large language models with performance close to full precision training [17].

D. Model Evaluation Module

The model was checked at the end of each training epoch to see how well it was learning. A custom function was used to measure its performance by calculating ROUGE scores to evaluate the quality of headlines. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures overlap between generated and reference headlines. The testing set evaluation helps in understanding the model's ability to generate accurate and coherent headlines from Nepali news articles.

- F1 Score:
  F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. A high F1 score means the model has a good balance of accuracy and completeness. The F1 score is calculated as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

- Precision:
  It measures the accuracy of positive predictions of a model. It shows how many of the predicted positives are correct. The Precision score is calculated as:

$$\text{Precision} = \frac{\text{Overlapping n-grams}}{\text{n-grams in candidate}} \tag{2}$$

- Recall:
  It measures how many of the actual positives were correctly identified. It shows the model's ability to find all positives. The Recall score is calculated as:

$$\text{Recall} = \frac{\text{Overlapping n-grams}}{\text{n-grams in reference}} \tag{3}$$

E. User Interface and API Module

- Backend: FastAPI is used to provide endpoints for real-time headline generation. It handles requests efficiently and ensures smooth communication between the frontend and the model.
- Frontend: It was developed using Flutter, the interface allows users to input news text and view generated headlines instantly. The mobile-friendly design ensures an intuitive and responsive user experience across devices.

IV. MODEL ARCHITECTURE

mBART is a sequence-to-sequence model for text production and comprehension developed by Facebook AI. It is an extension of the BART (Bidirectional and Auto-Regressive Transformers) model that supports multiple languages. This paradigm builds on the strengths of both BERT (Bidirectional Encoder Representations from Transformers) and GPT. The model is pre-trained on huge multilingual corpora, enabling it to perform tasks such as machine translation, summarization, and text synthesis in many languages. This multilingual feature allows mBART to transfer information from high-resource to low-resource languages, making it extremely useful for jobs involving multiple languages. It employs a denoising autoencoder architecture, with the model trained to reconstruct text from noisy input. This denoising objective assists the model in learning strong, generalized representations of language as well as improving its ability to handle missing or incomplete text data. The self-attention mechanism is used in both the encoder and the decoder of the mBART architecture. Self-attention in the encoder allows the model to capture long-term dependencies and context by processing the full input sequence at once. The decoder guarantees that the generated tokens are contextually meaningful by taking into account both the input and previously created tokens.

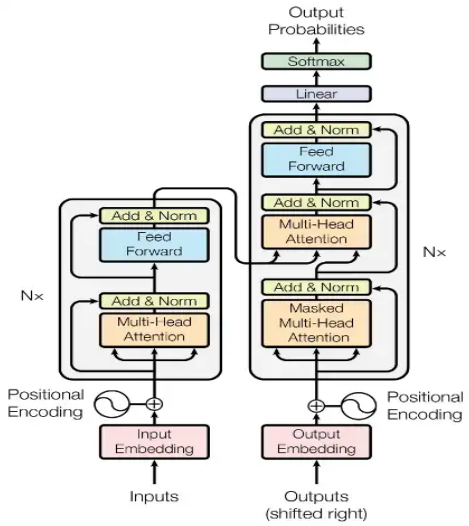

Figure 3: Architectute of Transformer Model [18]

A. Sequence-to-Sequence Probability

mBART models the probability of generating a headline $y = (y_1, y_2, \ldots, y_T)$ given an input news article $x = (x_1, x_2, \ldots, x_N)$ as:

$$P(y \mid x) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, x)$$

where $y_{<t}$ represents all previously generated tokens, and $P(y_t \mid y_{<t}, x)$ is computed by the decoder using the encoder's contextual embeddings.

B. Self-Attention (Scaled Dot-Product)

For each token, self-attention computes a weighted sum of all token embeddings in the input sequence:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

are the query, key, and value matrices, and d_k is the dimension of the key vectors.

The formal definition of how mBART creates a headline from a news story is given by the sequence-to-sequence probability equation. The transformer architecture, which employs self-attention processes to capture relationships between tokens, is used by the model to accomplish this. These methods are explained in depth in the sections that follow.

C. Encoder

The mBART encoder uses the BERT architecture to process text in both directions. Each of its six identical layers has two sub-layers: a position-wise feed-forward network and a multi-head self-attention mechanism. To stabilize training, residual connections and layer normalization are used around each sub-layer. It takes the input tokens and generates contextual embeddings that capture the relationships and dependencies between all tokens in the sequence. These embeddings provide a meaningful representation of the input text, which the decoder can then use to generate accurate output.

D. Decoder

The decoder generates text one token at a time in an autoregressive manner, much like GPT. It includes six similar levels as well, but each layer has three sub-layers: a feed-forward network, encoder-decoder cross-attention, and masked self-attention. Masking ensures that forecasts rely solely on already created tokens by preventing tokens from attending to future positions. Each token is produced based on the context of the input and the sequence of tokens already generated, ensuring coherent and contextually relevant output.

Table II: Encoder vs Decoder Comparison

Feature	Encoder	Decoder
Layers	6	6
Sub-layers	2	3
Attention	Self	Self + Cross
Masking	None	Yes
Purpose	Encode	Decode

V. RESULT AND DISCUSSION

A. Dataset Sourcing:

A CSV file containing URL, News Article, and Headlines from Nepali news website sources such as Online Khabar, BBC Nepali, and Ratopati with total of 86628 news articles. The categorized form of the news article is given below.

Table III: Data Statistics

S.N.	Category	Count
1	Education	2960
2	Opinion	2720
3	Health	3920
4	Entertainment	19160
5	Sports	22168
6	Economy	35360

B. Data Distribution:

The dataset is splitted into training(90%) and validation (10%). The *mBart50TokenizerFast* is used as tokenizer.

Table IV: Data Distibution in Training and Validation

Dataset Type	Count
Training Set	77965
Validation Set	8663

C. Model Training Setup:

Seq2SeqTrainingArguments is used to configure the training arguments. LoRA reduced the amount of trainable parameters while effectively fine-tuning the model. This method maintains excellent efficiency and accuracy while lowering memory utilization and expediting training without sacrificing model performance.

Table V: Training Parameters Summary

Parameter	Value
Model Name	facebook/mbart-large-50
Source Language	en_XX
Target Language	ne_NP
Max Length (Article)	512
Max Length (Headline)	50
Batch Size (Train)	8
Batch Size (Eval)	8
Learning Rate	0.0002
Num Train Epochs	3
Evaluation Strategy	epoch
Gradient Accumulation Steps	2
Warmup Steps	250
Quantization	4-bit
Model dtype	bfloat16
LoRA rank (r)	16
LoRA alpha	32
LoRA dropout	0.05

D. Visualization:

The training and validation loss are plotted using *matplotlib* to visualize the model's performance during training. These plots help in monitoring how well the model is learning over time, identifying issues such as overfitting or underfitting.

Table VI: Training and Validation Loss per Epoch

Epoch	Training Loss	Validation Loss
1	0.475100	0.454159
2	0.444300	0.441768
3	0.432600	0.437308

The above table shows model's performance changes during training, as the table displays the training loss and validation loss values over three epochs. Effective learning without overfitting is indicated by the steady decline in both training and validation loss.

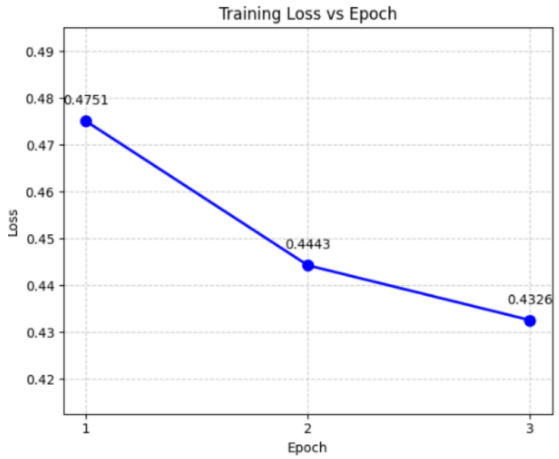

Figure 4: Graph Obtained from Training

The curve in Figure 2 illustrates training loss over epochs for a machine learning model. As the model learns and performs better in the training data, the training loss gradually drops from 0.47 to 0.43 throughout the epochs.

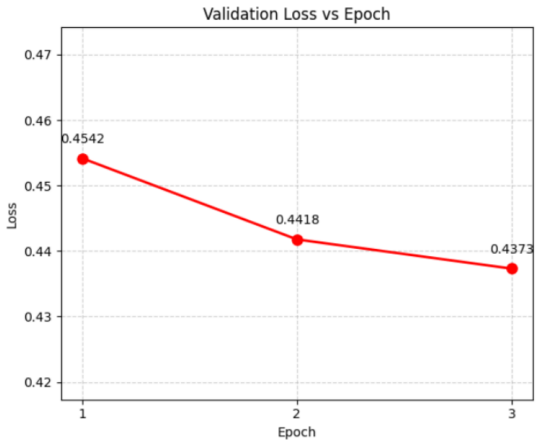

Figure 5: Graph Obtained from Validation

The curve in Figure 3 illustrates the validation loss over epochs for a machine learning model. The validation loss drops from 0.45 to about 0.43, indicating that the model is effectively generalizing to new data.

E. Output Generation:

The sample news articles from the validation set are tokenized, and the trained model generates Nepali headlines. The results are displayed alongside the original input and articles.

Table VII: A Sample of Manual and Machine-Generated Headlines for Nepali News Articles

News Article	Manual Headline	Machine Generated Headline
काठमाडौं । त्रिभुवन विश्वविद्यालयले स्वतन्त्र विद्यार्थी युनियन (स्ववियु) निर्वाचन चैत ५ गते हुने भएको छ । फागुन १५ गतेका लागि तोकिएको निर्वाचन चैत ५ गतेका लागि सरेको हो । त्रिवि विद्यार्थी कल्याण तथा खेलकुद निर्देशनालयका निर्देशक पशुपति अधिकारीले चैत ५ गते स्ववियु निर्वाचन गर्ने निर्णय गरिएको बताए ।	त्रिभुवन विश्वविद्यालयमा स्ववियु निर्वाचन सर्‍यो	स्ववियु निर्वाचन चैत ५ गते हुने

F. Model Performance Evaluation:

The model is evaluated based on ROUGE score. The score obtained is shown below.

Table VIII: Rouge Score for 1000 Samples

For 100 samples			
ROUGE	Recall	Precision	F1
Rouge 1	0.4736	0.4699	0.4682
Rouge 2	0.3037	0.3033	0.3014
Rouge L	0.4649	0.4671	0.4545

The model's ROUGE-1, F1 score of 0.4682 indicates good single-word overlap for the 100 sample dataset. With a ROUGE-2, F1 score of 0.3014, word pair capture is somewhat challenging. With a ROUGE-L, F1 score of 0.4545, the model shows good performance for a smaller dataset and is able to capture the longest common subsequence.

Table IX: Rouge Score for 2000 Samples

For 2000 samples			
ROUGE	Recall	Precision	F1
Rouge 1	0.4704	0.4372	0.4428
Rouge 2	0.2886	0.2625	0.2671
Rouge L	0.4465	0.4154	0.4205

With a ROUGE-1, F1 score of 0.4482 for the 2000 sample dataset, the model's performance is very steady and only slightly lowers single-word overlap. As the dataset size increases, the ROUGE-2, F1 score decreases to 0.2671, suggesting a minor difficulty with word pairs. Additionally, the ROUGE-L, F1 score decreases to 0.4205, indicating a minor decline in identifying the longest common subsequence.

Despite the larger dataset, the model performs well overall with minor variation.

VI. CONCLUSION

The system focused on enhancing the generation of Nepali news headline generation using multilingual model such as mBART, and incorporating techniques like LoRA and quantization. In the study, a diverse dataset of Nepali news article was collected and preprocessed to ensure its readiness for model training. The 4-bit quantised mBART with LoRA performed the best, generating accurate and cohesive headlines, according to evaluation using the ROUGE with score of 0.45.

This research paves the path for more effective and scalable news publishing by automating content creation in Nepali media. It guarantees consistency, lessens manual labor, and aids in the digital transformation of the Nepali news sector. This model's ability to generate high-quality native headlines has been astounding, demonstrating AI's potential in the Nepali language.

REFERENCES

[1] 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. [Online]. Available: https://arxiv.org/abs/2406.11289

[2] V. Karjule, J. Dange, S. Thange, J. Sase, and Prof. Kokate, "A Survey on Text Summarization Techniques," *Journal of Natural Language Processing*, Nov. 2023.

[3] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text Summarization Techniques: A Brief Survey," *arXiv preprint arXiv:1707.02268*, 2017. [Online]. Available: https://arxiv.org/abs/1707.02268

[4] Anonymous, "News Headline Generation Using Abstractive Text Summarization," *Proceedings of the International Conference on Artificial Intelligence and Data Science (ICAIDS)*, Springer, 2020, pp. 85–94, doi: 10.1007/978-3-031-20977-2_9.

[5] S. Babar, M. Tech-Cse, and Rit, "Text Summarization: An Overview," Oct. 2013.

[6] R. Khanal, S. Adhikari, and S. Thapa, "Extractive Method for Nepali Text Summarization Using Text Ranking and LSTM," in *Proc. 10th IOE Graduate Conf.*, May 2022.

[7] S. Patil, A. Pawar, S. Khanna, A. Tiwari, and S. Trivedi, "Text Summarizer Using NLP (Natural Language Processing)," *Computer Technology and Application*, vol. 12, pp. 20–21, Nov. 2022.

[8] K. Mishra, J. Rathi, and J. Banjara, "Encoder-Decoder Based Nepali News Headline Generation," *International Journal of Computer Applications*, vol. 175, pp. 975–8887, Sep. 2020, doi: 10.5120/ijca2020920735.

[9] P. Dhakal and D. Baral, "Abstractive Summarization of Low-Resourced Nepali Language Using Multilingual Transformers," *Unknown Journal*, Sep. 2024.

[10] A. Bukhtiyarov and I. Gusev, "Advances of Transformer-Based Models for News Headline Generation," arXiv preprint arXiv:2007.05044, 2020. [Online]. Available: https://arxiv.org/abs/2007.05044

[11] P. Li, J. Yu, J. Chen, and B. Guo, "HG-News: News Headline Generation Based on a Generative Pre-Training Model," *IEEE Access*, vol. PP, pp. 1–1, Aug. 2021, doi: 10.1109/ACCESS.2021.3102741.

[12] Z. Liu, G. I. Winata, and P. Fung, "Continual Mixed-Language Pre-Training for Extremely Low-Resource Neural Machine Translation," arXiv preprint arXiv:2105.03953, 2021. [Online]. Available: https://arxiv.org/abs/2105.03953

[13] B. Timalsina, N. Paudel, and T. B. Shahi, "Attention Based Recurrent Neural Network for Nepali Text Summarization," *Journal of Institute of Science and Technology*, vol. 27, pp. 141–148, Jun. 2022, doi: 10.3126/jist.v27i1.46709. The model achieved ROUGE-1: 0.46, ROUGE-2: 0.31, and ROUGE-L: 0.44, demonstrating improved coherence and informativeness over standard seq2seq models.

[14] S. Panthaplackel, A. Benton, and M. Dredze, "Updated Headline Generation: Creating Updated Summaries for Evolving News Stories," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pp. 6438–6461, Jan. 2022, doi: 10.18653/v1/2022.acl-long.446.

[15] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni, "Natural Language Processing (NLP) based Text Summarization - A Survey," in *Proc. 6th International Conference on Inventive Computation Technologies (ICICT 2021)*, pp. 1310–1317, Jan. 2021, doi: 10.1109/ICICT50816.2021.9358703.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021. https://arxiv.org/abs/2106.09685

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023. https://arxiv.org/abs/2305.14314

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv preprint arXiv:1706.03762*, Jun. 2023. [Online]. Available: https://arxiv.org/abs/1706.03762