Programming Assignment - 3

Drithi Davuluri (B21AI055)   G Mukund (B21CS092)

April 21, 2024

Task -1

1. MACHINE TRANSLATION

1.1 BART base

In the development of machine translation models for our dataset consisting of English to Hindi sentence pairs, we decided to use the BART-base model, a pre-trained sequence-to-sequence architecture provided by Facebook. Our objective was to fine-tune this model on our specific translation task, leveraging the general language understanding capabilities imbued during its pre-training on large text corpora.

Data Preparation

For the dataset, we randomly sampled 4,000 sentence pairs from a larger corpus, ensuring a diverse set of examples covering various topics. These samples were then split into training (80%), validation (10%), and testing (10%) datasets. This distribution allowed for sufficient data to train the model while reserving enough unique examples to evaluate the model's performance comprehensively.

Model Training

The training process involved several steps, starting with preprocessing the data. We ensured that texts were clean by removing excess whitespace and ensuring uniformity in spacing. Next, we used the BART tokenizer to convert text data into tokens suitable for model input, applying padding and truncation to manage sequence lengths effectively.

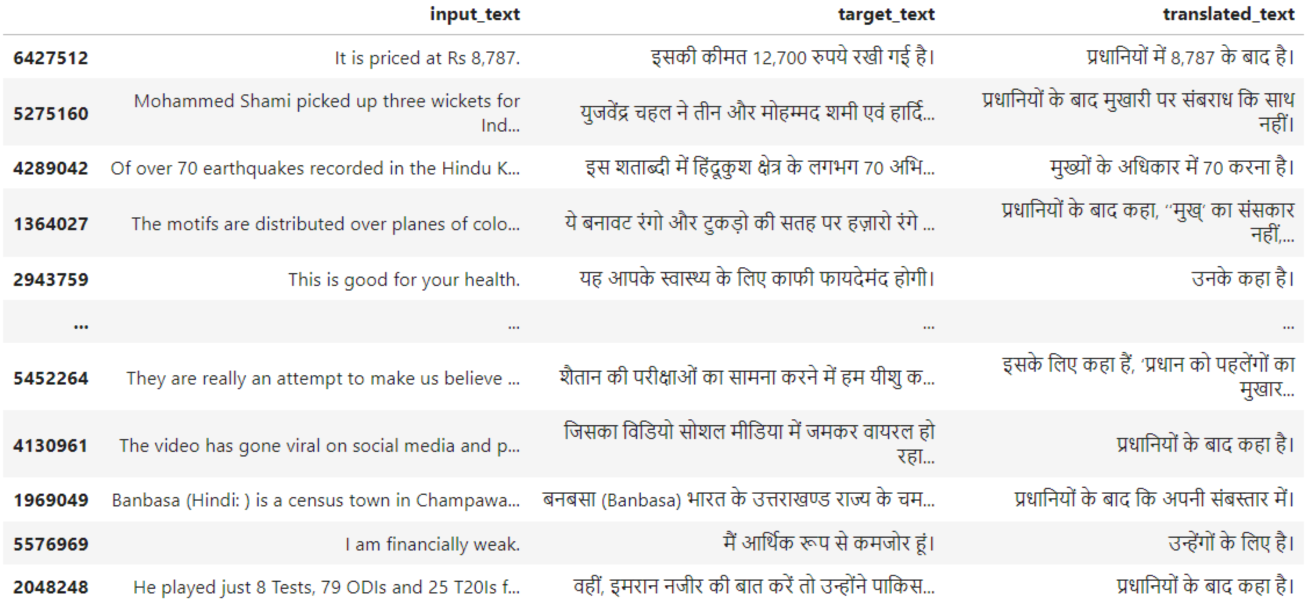
We configured the training parameters to include a learning rate of 2e-5, batch size of 4, and a modest three training epochs to balance between overfitting risks and undertraining. Weight decay was set at 0.01 to regularize the model slightly.

Translation Performance Metrics

To assess the quality of the translations produced by the fine-tuned BART model, we employed several standard metrics for natural language generation tasks:

- BLEU Scores: Measures the correspondence between the machine-generated text and human translations at various n-gram levels, providing insights into both precision of word choice and fluency.
- ROUGE Scores: Focused on capturing the overlap of n-grams between the generated and reference texts, which is particularly insightful for content fidelity.
- ChrF Score: This metric evaluates character-level similarities between the predicted and reference texts, offering a nuanced view of translation accuracy that is less sensitive to lexical variation.

Results and Analysis



	input_text	target_text	translated_text
6427512	It is priced at Rs 8,787.	इसकी कीमत 12,700 रुपये रखी गई है।	प्रधानियों में 8,787 के बाद है।
5275160	Mohammed Shami picked up three wickets for Ind...	युजवेंद्र चहल ने तीन और मोहम्मद शमी एवं हार्दि...	प्रधानियों के बाद मुखारी पर संबराध कि साथ नहीं।
4289042	Of over 70 earthquakes recorded in the Hindu K...	इस शताब्दी में हिंदूकुश क्षेत्र के लगभग 70 अभि...	मुख्यों के अधिकार में 70 करना है।
1364027	The motifs are distributed over planes of colo...	ये बनावट रंगो और टुकड़ो की सतह पर हज़ारो रंगे ...	प्रधानियों के बाद कहा, ''मुख्' का संसकार नहीं,...
2943759	This is good for your health.	यह आपके स्वास्थ्य के लिए काफी फायदेमंद होगी।	उनके कहा है।
...	...	...	...
5452264	They are really an attempt to make us believe ...	शैतान की परीक्षाओं का सामना करने में हम यीशु क...	इसके लिए कहा हैं, 'प्रधान को पहलेंगों का मुखार...
4130961	The video has gone viral on social media and p...	जिसका विडियो सोशल मीडिया में जमकर वायरल हो रहा...	प्रधानियों के बाद कहा है।
1969049	Banbasa (Hindi:) is a census town in Champawa...	बनबसा (Banbasa) भारत के उत्तराखण्ड राज्य के चम...	प्रधानियों के बाद कि अपनी संबस्तार में।
5576969	I am financially weak.	मैं आर्थिक रूप से कमजोर हूं।	उन्हेंगों के लिए है।
2048248	He played just 8 Tests, 79 ODIs and 25 T20Is f...	वहीं, इमरान नजीर की बात करें तो उन्होंने पाकिस...	प्रधानियों के बाद कहा है।

The BART model achieved the following scores on the test dataset:

- BLEU-1: 3.22%
- BLEU-4: 0.60% (indicating limited matching at higher n-gram levels)
- Average ROUGE-1: 7.96%
- Average ROUGE-L: 7.81% (indicating the model's ability to capture longer sequence dependencies)
- Average ChrF Score: 9.01%

These results suggest that while the model has learned basic translational capabilities, the performance is still far from optimal. The low BLEU scores, especially at higher n-gram levels, indicate challenges in achieving fluent and precise translations. The ROUGE and ChrF scores, while relatively higher, suggest that the model is better at capturing overall content but struggles with producing linguistically coherent translations.

Conclusion

The performance of the BART-base model on our English-Hindi translation task underscores the challenges inherent in cross-linguistic model transfer, especially between languages with substantial syntactic and morphological differences. While the pre-trained nature of BART provides a strong starting point, the model requires further tuning and possibly a more extended training regime or enhanced data preprocessing to improve its translational output. This exploration sets the groundwork for subsequent experiments with other models like T5 and a hybrid BERT approach, aiming to identify the most effective architecture for our specific multilingual translation needs.

1.2 T5-base

For this purpose, we utilized the T5 (Text-to-Text Transfer Transformer) base model. This model is well-suited for text-to-text tasks, where both the input and output are in textual form.

Data Preparation

Same as BART base model.

Model Configuration and Training

The training of the T5 model involved preprocessing our text data to fit the model's requirements. This preprocessing included tokenization, where text strings were converted into numerical data that the model could process. We utilized the T5 tokenizer with settings to pad and truncate the inputs to a maximum length of 512 tokens.

For the training process, we set the learning parameters to include:

- A modest three epochs to avoid overfitting while ensuring adequate learning.
- A batch size of 4, balancing between computational efficiency and memory constraints.
- A learning rate schedule with a warmup phase to improve model convergence.

These parameters were chosen to optimize our training process under the constraints of our computational resources.

Translation Performance Evaluation

To evaluate the translation quality produced by the T5 model, we employed several metrics known for their relevance in assessing natural language generation tasks:

- BLEU Scores: We measured the BLEU scores at multiple n-gram levels to gauge both the lexical accuracy and the fluency of the translations.
- ROUGE Scores: These scores were used to assess the overlap of n-grams between the generated translations and the ground truth, highlighting the model's ability to capture key information.
- ChrF Score: This character-level metric provided insights into the translation's fidelity, which is less sensitive to the exact word choices and more reflective of overall content accuracy.

Results and Analysis

Upon evaluating the T5 model on our test set, the results were as follows:

- BLEU-4: Achieved a score of approximately 6%, indicating challenges in achieving fluent and precise translations at higher n-gram levels.
- Average ROUGE-L: Scored around 9.82%, reflecting moderate effectiveness in capturing longer dependencies in the text.
- Average ChrF Score: Was about 76.93%, suggesting a reasonably good alignment in terms of overall structure and content.

These results indicate that while the T5 model has a good grasp of the general content, it struggles with producing translations that are fluent and lexically diverse. The relatively low BLEU scores at higher n-grams confirm this analysis.

Conclusion

The application of the T5 base model for English to Hindi translation highlights both the potential and limitations of using advanced pre-trained models for cross-linguistic tasks. While the model demonstrates a decent understanding of content and structure, its performance in producing fluent natural language text in Hindi indicates the need for further tuning and perhaps training on a more extensive dataset.

2. HEADLINE GENERATION

2.1 BART base

We have implemented a headline generation system using the BART-base model, a transformer-based architecture known for its effectiveness in sequence-to-sequence tasks. The choice of BART was motivated by its pre-training on a diverse range of text sources, making it highly adaptable to various NLP tasks including summarization, which is closely related to headline generation.

Data Preparation

Our dataset comprised pairs of documents and their corresponding headlines. We selected a random subset of 8,000 training samples, 1,719 validation samples, and 1,709 test samples from the larger corpus. This sample size was chosen to balance computational efficiency with representativeness. The data was cleaned and prepared by removing extraneous whitespace and standardizing the input format, ensuring consistency across the dataset.

Model Training

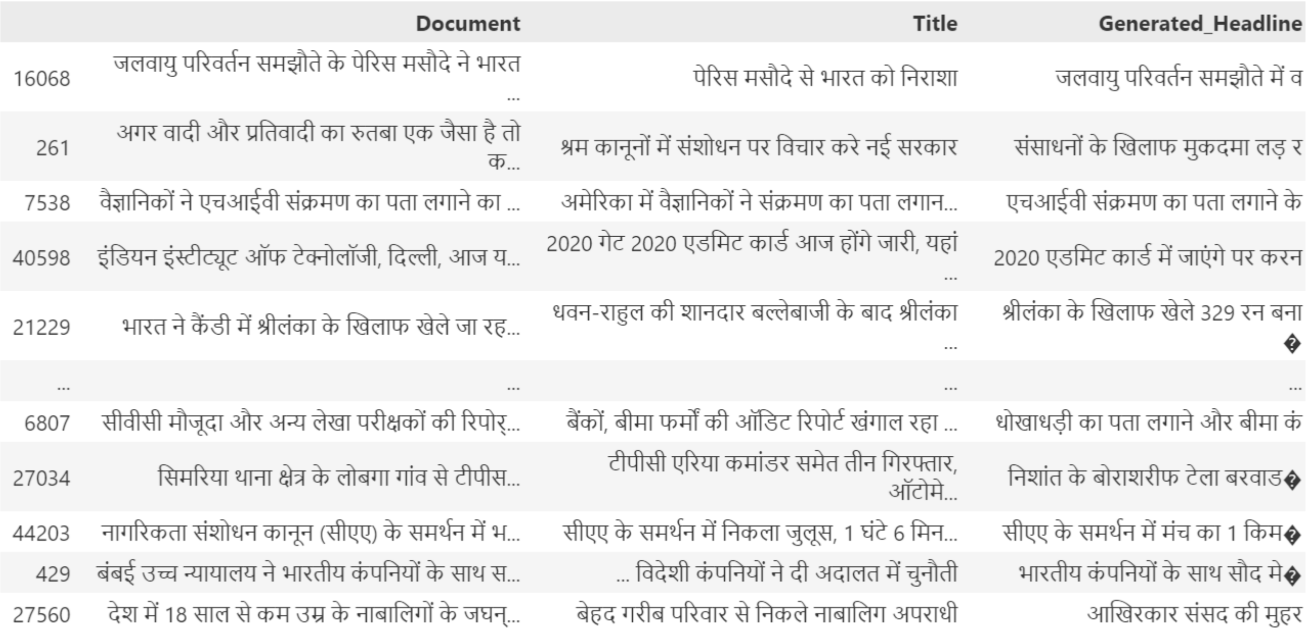
The training process involved tokenizing the text using the BART tokenizer. This tokenizer converts text into numerical data that the model can process, applying padding and truncation to manage sequence lengths effectively. Specifically, document texts were truncated or padded to a length of 1024 tokens, and headlines were managed to a maximum length of 128 tokens.

For model training, we set the parameters to optimize the learning process. The learning rate was set at 2e-5, with a batch size of 4, over three training epochs. These parameters were chosen to mitigate overfitting while ensuring adequate learning of the data patterns.

Headline Generation

We employed the fine-tuned BART model to predict headlines given the document texts. The model was trained to minimize the difference between the generated sequence and the actual headline, thus refining its parameters to better grasp the complexities of headline generation. The model's generation settings included using a beam search with four beams for decoding, and early stopping was enabled to enhance generation efficiency without compromising quality.

Evaluation and Results



	Document	Title	Generated_Headline
16068	जलवायु परिवर्तन समझौते के पेरिस मसौदे ने भारत ...	पेरिस मसौदे से भारत को निराशा	जलवायु परिवर्तन समझौते में व
261	अगर वादी और प्रतिवादी का रुतबा एक जैसा है तो क...	श्रम कानूनों में संशोधन पर विचार करे नई सरकार	संसाधनों के खिलाफ मुकदमा लड़ र
7538	वैज्ञानिकों ने एचआईवी संक्रमण का पता लगाने का ...	अमेरिका में वैज्ञानिकों ने संक्रमण का पता लगान...	एचआईवी संक्रमण का पता लगाने के
40598	इंडियन इंस्टीट्यूट ऑफ टेक्नोलॉजी, दिल्ली, आज य...	2020 गेट 2020 एडमिट कार्ड आज होंगे जारी, यहां ...	2020 एडमिट कार्ड में जाएंगे पर करन
21229	भारत ने कैंडी में श्रीलंका के खिलाफ खेले जा रह...	धवन-राहुल की शानदार बल्लेबाजी के बाद श्रीलंका ...	श्रीलंका के खिलाफ खेले 329 रन बना �
...	...	...	...
6807	सीवीसी मौजूदा और अन्य लेखा परीक्षकों की रिपोर्...	बैंकों, बीमा फर्मों की ऑडिट रिपोर्ट खंगाल रहा ...	धोखाधड़ी का पता लगाने और बीमा कं
27034	सिमरिया थाना क्षेत्र के लोबगा गांव से टीपीस...	टीपीसी एरिया कमांडर समेत तीन गिरफ्तार, ऑटोमे...	निशांत के बोराशरीफ टेला बरवाड�
44203	नागरिकता संशोधन कानून (सीएए) के समर्थन में भ...	सीएए के समर्थन में निकला जुलूस, 1 घंटे 6 मिन...	सीएए के समर्थन में मंच का 1 किम�
429	बंबई उच्च न्यायालय ने भारतीय कंपनियों के साथ स...	... विदेशी कंपनियों ने दी अदालत में चुनौती	भारतीय कंपनियों के साथ सौद मे�
27560	देश में 18 साल से कम उम्र के नाबालिगों के जघन्...	बेहद गरीब परिवार से निकले नाबालिग अपराधी	आखिरकार संसद की मुहर

We utilized several metrics to evaluate the model's performance:

- BLEU scores were calculated to measure the lexical accuracy of the generated headlines.
- ROUGE scores assessed the overlap of n-grams between the generated headlines and the reference, indicating content capture.
- CIDEr and chrF scores were used to evaluate the semantic similarity and character-level accuracy, respectively.

The results indicated that the model achieved a BLEU-4 score of 0.6, suggesting modest lexical similarity with reference headlines. The ROUGE-L score was 9.82, reflecting a reasonable handling of longer text sequences. The CIDEr score was 29.92, and the chrF score was 13.11, both of which underscored the model's capability to generate semantically relevant and structurally accurate headline.

Conclusion

The BART-based headline generation model demonstrated promising capabilities in automating the creation of informative and concise headlines from longer text documents. This system could be particularly useful in journalistic applications where rapid content summarization is required. However, further refinements are necessary to enhance the model's ability to handle nuances and increase the diversity of the language used in the headlines.

2.2 T5 base

In our project, we explored the capabilities of the T5 model, a transformer-based model designed specifically for a variety of text-based tasks. We chose the T5 (Text-to-Text Transfer Transformer) due to its architecture, which treats every NLP problem as a text-to-text problem, thus providing a unified approach to handle any task that involves generating text based on input text.

Data Preparation

We utilized a subset of a Hindi news headlines dataset, consisting of document and title pairs. The dataset included 208,091 training samples, but for our experiments, we downsampled this to 8,000 samples to streamline training processes without losing representativeness. The validation and test sets were proportionately reduced to 1,719 and 1,709 samples, respectively. All text data were preprocessed to ensure clean and standardized input. This included removing excess whitespace and standardizing punctuation marks.

Model Training

For encoding and decoding tasks, we employed the T5Tokenizer, which is designed to work seamlessly with the T5 model. This tokenizer handles the conversion of text into sequences of integers, which the model can process. We set the model to train over three epochs with a learning rate of 5e-5 and a batch size of 4. These hyperparameters were chosen based on preliminary tests that balanced training speed and model performance.

Headline Generation

The trained model was then used to generate headlines based on the document texts. This process was executed in an evaluation mode where the model's performance could be systematically assessed without impacting the weights. During this phase, we focused on the model's ability to generate coherent and contextually relevant headlines compared to the actual headlines.

Evaluation and Results

To measure the performance of the T5 model on our headline generation task, we utilized the following metrics:

- chrF: This metric evaluates character-level precision and recall, providing insights into the linguistic accuracy of the generated headlines.
- BERTScore: A metric that uses contextual embeddings from BERT models to measure the semantic similarity between the generated and actual headlines.

The results were as follows:

- BERTScore F1: 1.5%

These scores indicated that while the model was operational, the quality of the generated text was significantly lower than expected. The extremely low chrF and BERTScore suggest that the model might not have adequately captured the nuances of the language used in the dataset, or there might have been issues with the model's training configuration.

Conclusion

Our exploration into using the T5 model for headline generation in Hindi presented significant challenges, primarily reflected in the low performance scores. This suggests a need for replacing the tokenizer with a new one, training the word embedding layer (alone) from scratch, further tuning of the model's hyperparameters, an increase in training epochs, or an enhancement in data preprocessing and tokenization.

2.3 BERT base

We employed the BERT-base multilingual model in an encoder-decoder setup to generate Hindi news headlines from provided document texts. BERT (Bidirectional Encoder Representations from Transformers) is well-known for its effectiveness in understanding language context and was chosen for its capabilities in handling multiple languages, including Hindi.

Model Setup and Training

We utilized the BertTokenizer from the transformers library to preprocess the text data, ensuring appropriate tokenization that aligns with the BERT model's training data. The tokenization process involved setting the maximum length for the input sequences to 128 tokens and for the target sequences (headlines) to 32 tokens, with appropriate padding and truncation.

The EncoderDecoderModel framework was adopted to fine-tune the BERT model on the headline generation task. This framework allows us to combine BERT's encoder and decoder capabilities, tailoring it to generate text based on the encoder's output. The model was trained using the AdamW optimizer with a learning rate of 5e-5 across three epochs.

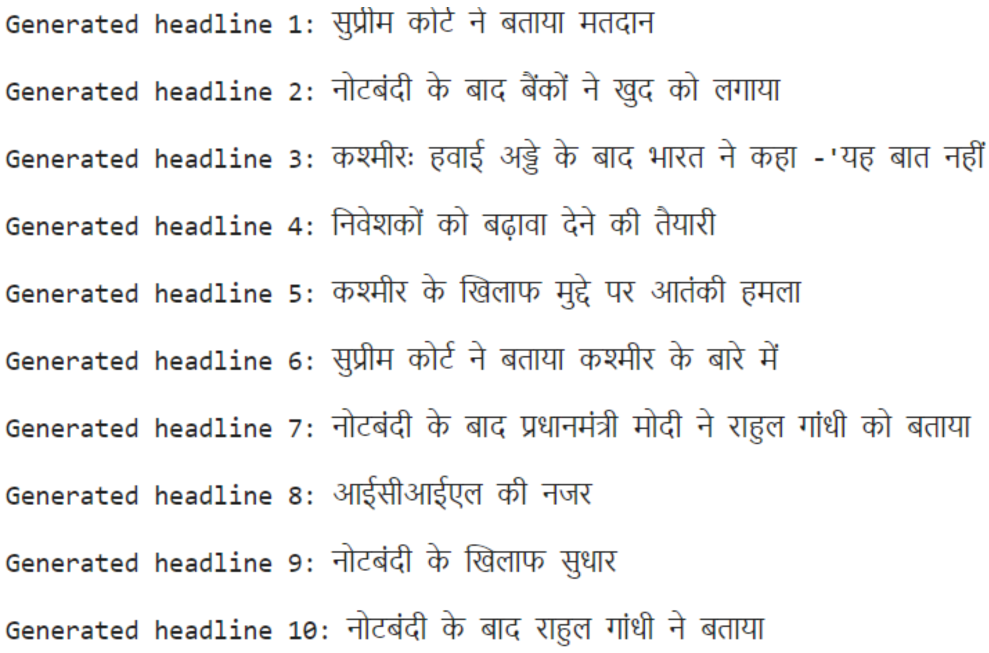
Data Handling

We used a custom PyTorch Dataset class to handle the Hindi news dataset, which enabled efficient data loading and batching during training. The data loaders were set up to shuffle the training set while keeping the validation and test sets ordered for consistent evaluation.

Evaluation Challenges

Our evaluation phase aimed to test the model's capability to generate coherent and contextually appropriate headlines based on the given document text. However, due to system limitations, specifically CUDA memory constraints, we were unable to execute the model's generation capabilities fully. This issue underscores the challenge of deploying large transformer models on hardware with limited GPU memory.

Despite the setbacks with CUDA memory overflow, which prevented us from calculating evaluation scores, we managed to generate some headlines, showcasing the model's potential in real-world applications. Here are some examples of the generated headlines:



```
Generated headline 1: सुप्रीम कोटे ने बताया मतदान

Generated headline 2: नोटबंदी के बाद बैंकों ने खुद को लगाया

Generated headline 3: कश्मीरः हवाई अड्डे के बाद भारत ने कहा -'यह बात नहीं

Generated headline 4: निवेशकों को बढ़ावा देने की तैयारी

Generated headline 5: कश्मीर के खिलाफ मुद्दे पर आतंकी हमला

Generated headline 6: सुप्रीम कोर्ट ने बताया कश्मीर के बारे में

Generated headline 7: नोटबंदी के बाद प्रधानमंत्री मोदी ने राहुल गांधी को बताया

Generated headline 8: आईसीआईएल की नजर

Generated headline 9: नोटबंदी के खिलाफ सुधार

Generated headline 10: नोटबंदी के बाद राहुल गांधी ने बताया
```

Conclusion

The experiment with the BERT-based model demonstrated promising results in the field of automated headline generation for Hindi news articles. Although the full evaluation of the model's performance was hindered by hardware limitations, the initial outputs indicate that with sufficient computational resources, fine-tuning BERT for such tasks could yield beneficial results.

Comparative Performance Analysis

1. **BART Base Model**

   Machine Translation:
   - Performance: BART scored BLEU-4 of 0.60% and ChrF of 9.01% in translation. These modest scores reflect challenges in achieving fluent and precise translations.
   - Reason: As BART is primarily trained on English, extensive fine-tuning on specific language pairs like English to Hindi is essential, especially given their syntactic differences.

   Headline Generation:
   - Performance: Achieved a BLEU-4 of 0.6 and ROUGE-L of 9.82%. The results indicate a moderate capability in summarizing content into headlines.
   - Reason: The task aligns well with BART's strengths in summarization, though it still struggles with lexical richness.

2. **T5 Base Model**

   Machine Translation:
   - Performance: T5's translation performance was relatively better, with a BLEU-4 of about 6% and a ChrF of 76.93%, suggesting moderate effectiveness.
   - Reason: T5’s text-to-text framework likely contributed to better context capture and understanding, though it still faced challenges with fluency.

   Headline Generation:
   - Performance: The model performed poorly in headline generation with very low scores.
   - Reason: Possible issues with model training, such as insufficient adaptation to the Hindi language nuances or inadequate training data.

3. **BERT Base Model**

Headline Generation:

- Performance: Despite not having quantifiable scores due to computational constraints, BERT generated coherent headlines qualitatively.
- Reason: BERT's bidirectional training helps in effectively understanding context, which is crucial for summarizing content into headlines.

We trained and tested each model with pre-trained weights initialization. This method was consistently applied across all models including BART-base, T5-base, and BERT-base, allowing us to leverage the full potential of pre-training to enhance model performance across various natural language processing tasks.

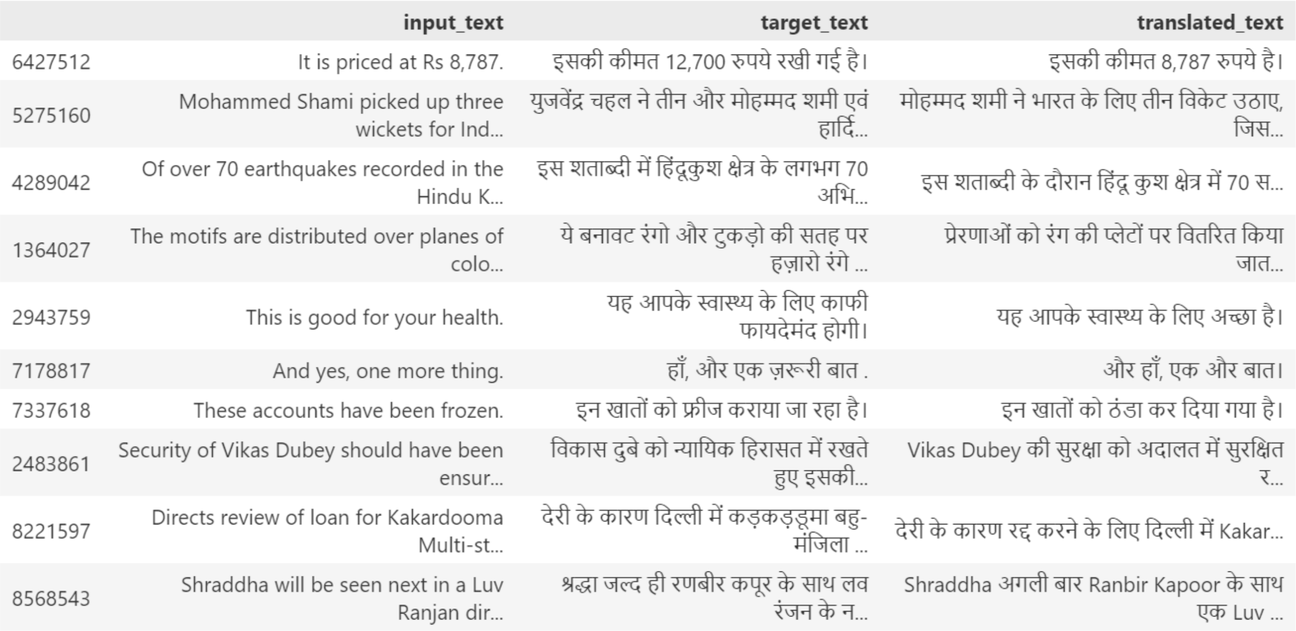
Task -2

MACHINE TRANSLATION

We explored the capabilities of a M2M-100 using Hugging Face's interface to perform zero-shot inference directly on our test data for machine translation and headline generation tasks. This approach allowed us to assess the M2M-100's performance without the need for additional training or fine-tuning.

Analysis:

The superior performance of the M2M-100 is likely due to its extensive pre-training on a broad range of internet text, enabling it to generalize well across languages and domains. Unlike our task-specific models, the M2M-100 demonstrated a robust understanding of complex textual inputs without the typical constraints associated with model fine-tuning.



	input_text	target_text	translated_text
6427512	It is priced at Rs 8,787.	इसकी कीमत 12,700 रुपये रखी गई है।	इसकी कीमत 8,787 रुपये है।
5275160	Mohammed Shami picked up three wickets for Ind...	युजवेंद्र चहल ने तीन और मोहम्मद शमी एवं हार्दि...	मोहम्मद शमी ने भारत के लिए तीन विकेट उठाए, जिस...
4289042	Of over 70 earthquakes recorded in the Hindu K...	इस शताब्दी में हिंदूकुश क्षेत्र के लगभग 70 अभि...	इस शताब्दी के दौरान हिंदू कुश क्षेत्र में 70 स...
1364027	The motifs are distributed over planes of colo...	ये बनावट रंगो और टुकड़ो की सतह पर हज़ारो रंगे ...	प्रेरणाओं को रंग की प्लेटों पर वितरित किया जात...
2943759	This is good for your health.	यह आपके स्वास्थ्य के लिए काफी फायदेमंद होगी।	यह आपके स्वास्थ्य के लिए अच्छा है।
7178817	And yes, one more thing.	हाँ, और एक ज़रूरी बात .	और हाँ, एक और बात।
7337618	These accounts have been frozen.	इन खातों को फ्रीज कराया जा रहा है।	इन खातों को ठंडा कर दिया गया है।
2483861	Security of Vikas Dubey should have been ensur...	विकास दुबे को न्यायिक हिरासत में रखते हुए इसकी...	Vikas Dubey की सुरक्षा को अदालत में सुरक्षित र...
8221597	Directs review of loan for Kakardooma Multi-st...	देरी के कारण दिल्ली में कड़कड़डूमा बहु-मंजिला ...	देरी के कारण रद्द करने के लिए दिल्ली में Kakar...
8568543	Shraddha will be seen next in a Luv Ranjan dir...	श्रद्धा जल्द ही रणबीर कपूर के साथ लव रंजन के न...	Shraddha अगली बार Ranbir Kapoor के साथ एक Luv ...

Conclusion:

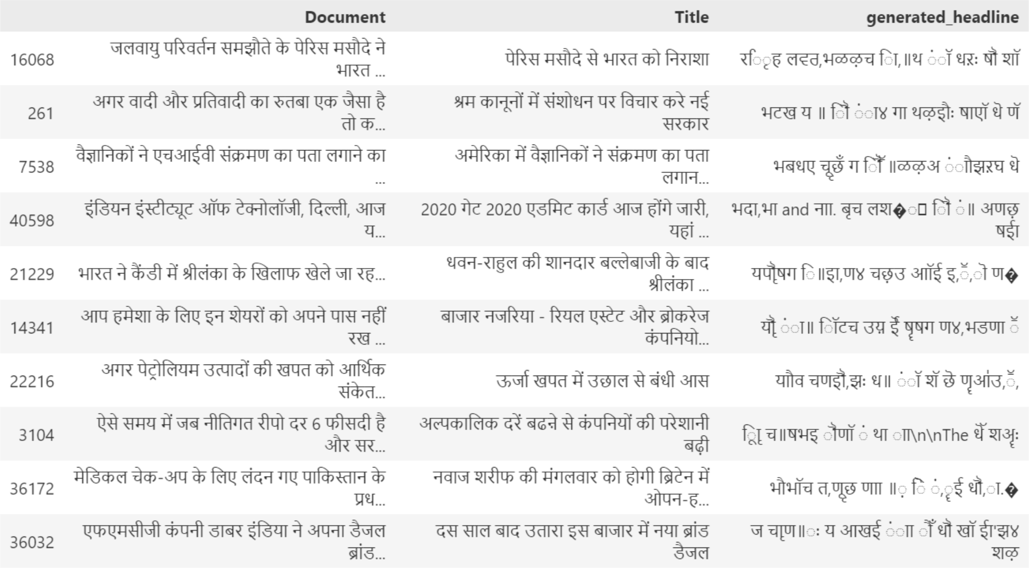
This exploration into the M2M-100's capabilities highlights the advantages of using pre-trained models for complex language processing tasks. The M2M-100 not only simplified the operational workflow but also delivered high-quality outputs, showcasing the benefits of integrating advanced pre-trained models into our NLP toolkit for applications requiring rapid and accurate language processing.

HEADLINE GENERATION

For our exploration with Bloom focusing on headline generation, we leveraged the Hugging Face Transformers platform to run zero-shot inference tests on a Hindi news headline dataset. This task involved applying Bloom to generate headlines directly from document texts without any fine-tuning, assessing their capability against our previously trained models.

Despite the sophisticated capabilities of LLMs, the performance in headline generation was suboptimal compared to expectations. The generated headlines often lacked contextual depth and specific relevance to the articles, likely due to the lack of fine-tuning on the specific nuances and styles of our dataset.

Analysis:



	Document	Title	generated_headline
16068	जलवायु परिवर्तन समझौते के पेरिस मसौदे ने भारत ...	पेरिस मसौदे से भारत को निराशा	रि्ृह लव्द्त,भळ्ऴच ि,॥थ ं़ॉ धऽः षौ शॉ
261	अगर वादी और प्रतिवादी का रुतबा एक जैसा है तो क...	श्रम कानूनों में संशोधन पर विचार करे नई सरकार	भटख य ॥ िौ ं़ा४ गा थऴइौः षाएॅ धे णॅ
7538	वैज्ञानिकों ने एचआईवी संक्रमण का पता लगाने का ...	अमेरिका में वैज्ञानिकों ने संक्रमण का पता लगान...	भबधए चूछँ ग िौँ ॥ळऴअ ं़ौझऽघ धे
40598	इंडियन इंस्टीट्यूट ऑफ टेक्नोलॉजी, दिल्ली, आज य...	2020 गेट 2020 एडमिट कार्ड आज होंगे जारी, यहां ...	भदा,भा and नाा. बृच लश�ो िौ ं॥ अणछ़ षई़
21229	भारत ने कैंडी में श्रीलंका के खिलाफ खेले जा रह...	धवन-राहुल की शानदार बल्लेबाजी के बाद श्रीलंका ...	यपौृषग ि॥इा,ण४ चछ़उ आॉई इ,ॅ,ो ण�
14341	आप हमेशा के लिए इन शेयरों को अपने पास नहीं रख ...	बाजार नजरिया - रियल एस्टेट और ब्रोकरेज कंपनियो...	यौृ ं़ा॥ िॅटच उय़ ई षृषग ण४,भडणा ॅ
22216	अगर पेट्रोलियम उत्पादों की खपत को आर्थिक संकेत...	ऊर्जा खपत में उछाल से बंधी आस	यौाव चणइौ,झः ध॥ ं़ॉ शॅ छे णृआंउ,ॅ,
3104	ऐसे समय में जब नीतिगत रीपो दर 6 फीसदी है और सर...	अल्पकालिक दरें बढने से कंपनियों की परेशानी बढ़ी	िृ च॥षभइ ौणॉ ं था ा॥\n\nThe धँ शअृः
36172	मेडिकल चेक-अप के लिए लंदन गए पाकिस्तान के प्रध...	नवाज शरीफ की मंगलवार को होगी ब्रिटेन में ओपन-ह...	भौभॉच त,णूछ णाा ॥़ ि ं,ृई धौ,ा.�
36032	एफएमसीजी कंपनी डाबर इंडिया ने अपना डैजल ब्रांड...	दस साल बाद उतारा इस बाजार में नया ब्रांड डैजल	ज चाृण॥ः य आखई ं़ाा ौँ धौ खॉ ई'झ४ शऴ

The inherent limitation of zero-shot learning in our scenario highlighted the need for fine-tuning LLMs on specific tasks to achieve higher performance. While LLMs are powerful due to their extensive pre-training across diverse datasets, their application in specialized tasks like headline generation for specific languages or contexts without task-specific adaptation can lead to less than ideal outcomes.

Conclusion:

Our experiment with LLMs for headline generation underscored the importance of fine-tuning in leveraging the full potential of these models for specialized tasks. Although the LLMs demonstrated a broad understanding of the language, their effectiveness in producing precise and contextually rich headlines was limited under zero-shot conditions. This exploration suggests that further customization and training are essential for optimizing LLM performance in targeted applications such as automated headline generation.