PRESENTATION ON

“Emphasizing Seq2Seq architecture using T5 and BART models in Question Answering”

Submitted in partial fulfillment of the requirements of the degree of

Bachelor Of Technology

in

Modern Machine Learning

For the academic year 2024-2025

Contents

- Introduction
- Literature Review
- Comparison Table
- Limitations
- Objectives
- Methodology
- Implementation Workflow
- Result and Conclusion
- References

INTRODUCTION

- Seq2Seq (sequence-to-sequence) models, introduced by Google in 2014, have transformed NLP by handling variable-length input and output tasks.
- Based on an encoder-decoder architecture, Seq2Seq models capture essential information from input sequences and generate coherent outputs.
- These models excel in tasks like question answering and text summarization, making them widely applicable.
- The attention mechanism enhances Seq2Seq by focusing on relevant parts of the input, improving accuracy, especially for long sequences.
- Seq2Seq applications span machine translation, chatbots, and summarization, providing significant economic and operational benefits.
- Ongoing advancements promise to expand the capabilities and impact of Seq2Seq models across NLP domains.

Literature Survey

Ref	Journal	Publish held in	Title	Methodology
Jacob Devlin et al.	NAACL	2019	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Bidirectional Transformer model using a multi-layer encoder with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) for deep contextual representation.
Colin Raffel et al.	Journal of Machine Learning Research	2020	T5: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	Unified text-to-text framework; pre-trained on large dataset (C4) with denoising autoencoder objective, fine-tuned on task-specific data.
Kevin Clark et al.	ICLR	2020	ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators	Introduces replaced token detection task with generator-discriminator model, providing efficient learning through detection of altered tokens.

Ref	Journal	Publish held in	Title	Methodology
Zhenzhong Lan et al.	ICLR	2019	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations	Parameter reduction via factorized embedding and cross-layer parameter sharing; introduces Sentence-Order Prediction (SOP) for inter-sentence coherence.
Mike Lewis et al.	ACL	2019	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Processing	Combines bidirectional encoder and autoregressive decoder, trained as a denoising autoencoder using various noising functions like token deletion and span masking.

Comparison Table of different Models

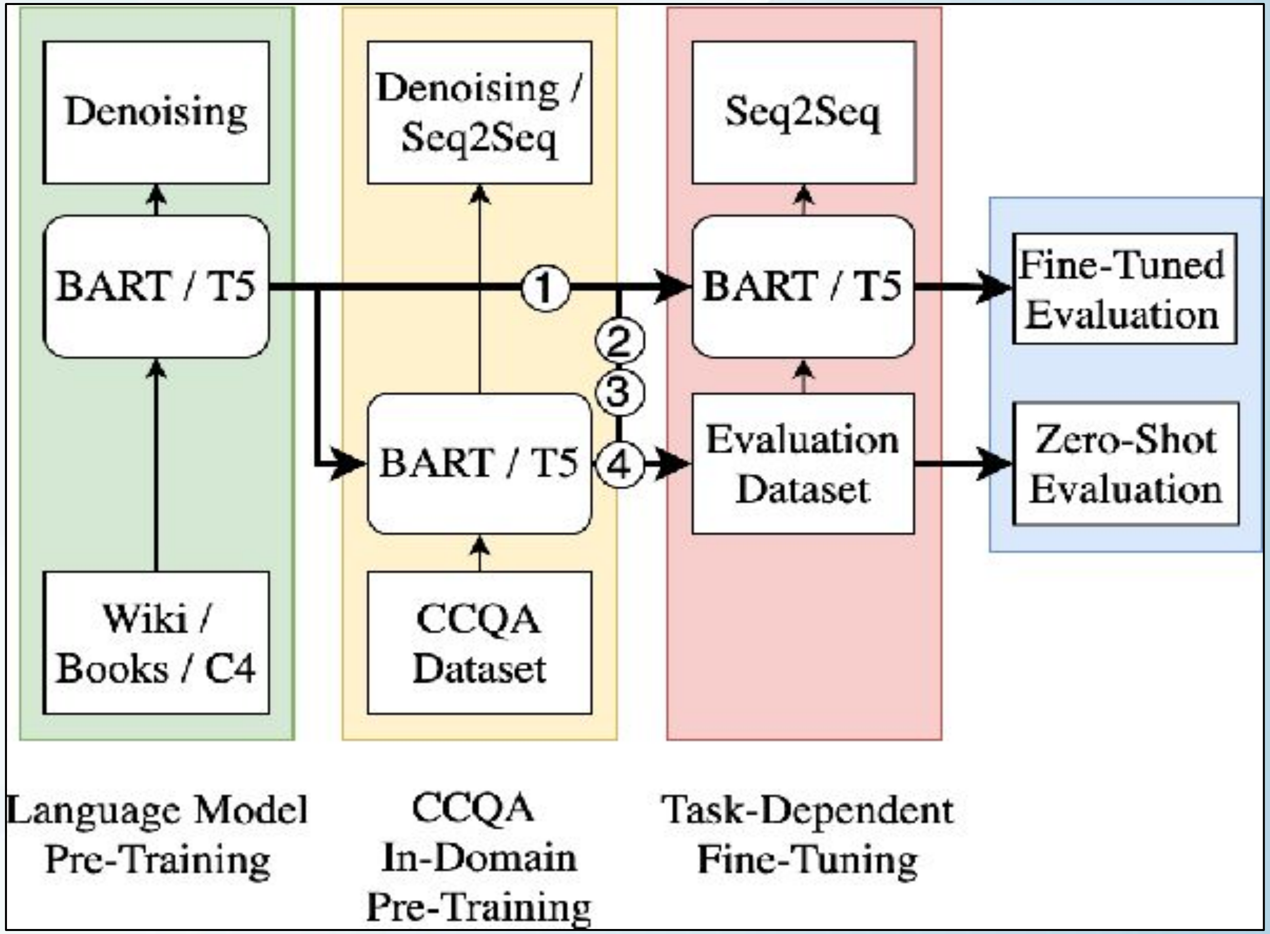
Model	Key Feature
BERT	- **Bidirectional Context** with Masked Language Modeling - **NSP** for sentence-pair tasks
ELECTRA	- **Replaced Token Detection** (efficient training) - **Generator-Discriminator** model
T5	- **Text-to-Text Framework** for all NLP tasks - **Span Masking** on large web data
BART	- **Seq2Seq Architecture** (BERT-style encoder + GPT-style decoder) - **Denoising** for text generation
ALBERT	**-Parameter-Efficient** (factorized embeddings, cross-layer sharing) - **SOP** for sentence order.

MODEL	LIMITATIONS
BERT	- Requires substantial computational resources and memory, making it less accessible for smaller systems. - Pre-trained on masked language modeling, which may not capture the full context of sentences effectively.
DistilBERT	- Although it is smaller and faster, it may sacrifice some accuracy compared to the full BERT model. - May not perform as well on tasks that require deeper contextual understanding due to its reduced architecture
ALBERT	- While it reduces parameters through factorized embeddings, it may lead to a loss of expressiveness compared to BERT. - The training process can be more complex, potentially leading to longer training times.
RoBERTa	- Requires extensive computational resources for training due to the larger batch sizes and longer training times. - Lacks the ability to handle out-of-vocabulary words effectively, which may limit its performance on certain tasks.

Objective

- To develop an efficient question answering system that provides concise and meaningful answer of of any context based question.
- To implement the BART (Bidirectional and Auto-Regressive Transformers) and T5 (Text-To-Text Transfer Transformer) models, exploring their strengths in natural language understanding and generation.

- To evaluate the **advantages of using BART and T5 for summarization**, including:
  - **High Accuracy:** Both models achieve state-of-the-art performance in generating coherent and contextually rich summaries.
  - **Versatility:** BART and T5 can handle multiple NLP tasks, offering flexibility and adaptability across various text genres.
  - **Pre-trained on Large Corpora:** Enhances summary quality by leveraging a diverse knowledge base.
  - **Fine-tuning Capabilities:** Easy to customize for specific summarization needs, enabling domain-specific applications.



Methodology

1.Data Preparation

- Selected the Stanford Question Answering Dataset (SQuAD) for training.
- **About SQuAD:** Contains context-question-answer triplets from Wikipedia for supervised training.
- **Preprocessing:**
  - Tokenized context and question inputs using T5Tokenizer and BartTokenizer.
  - Concatenated and tokenized context and question, separated answer span as the label.
  - Converted tokens to PyTorch tensors with input_ids, attention_mask, and labels.
- **Edge Cases:** Handled empty answers and consolidated multiple answers; imposed a 128-token limit for efficiency.

2.Model Selection

- **T5 Small:** Text-to-text format, generates answers flexibly, low computational demand.
- **BART Base:** Suitable for text generation, combines bidirectional encoding and autoregressive decoding.

3.Model Training and Fine-Tuning

- **Training Configuration:**
  - **Epochs: 3, Batch Size: 8, Learning Rate: 3e-5.**
  - **Gradient Accumulation:** 4 steps, FP16 precision for BART to enhance speed and memory efficiency.
- **Setup:**
  - Used **Hugging Face Trainer API** for simplified training, with logging and checkpointing.
  - Recorded accuracy, F1-score, and loss to track convergence and detect overfitting.

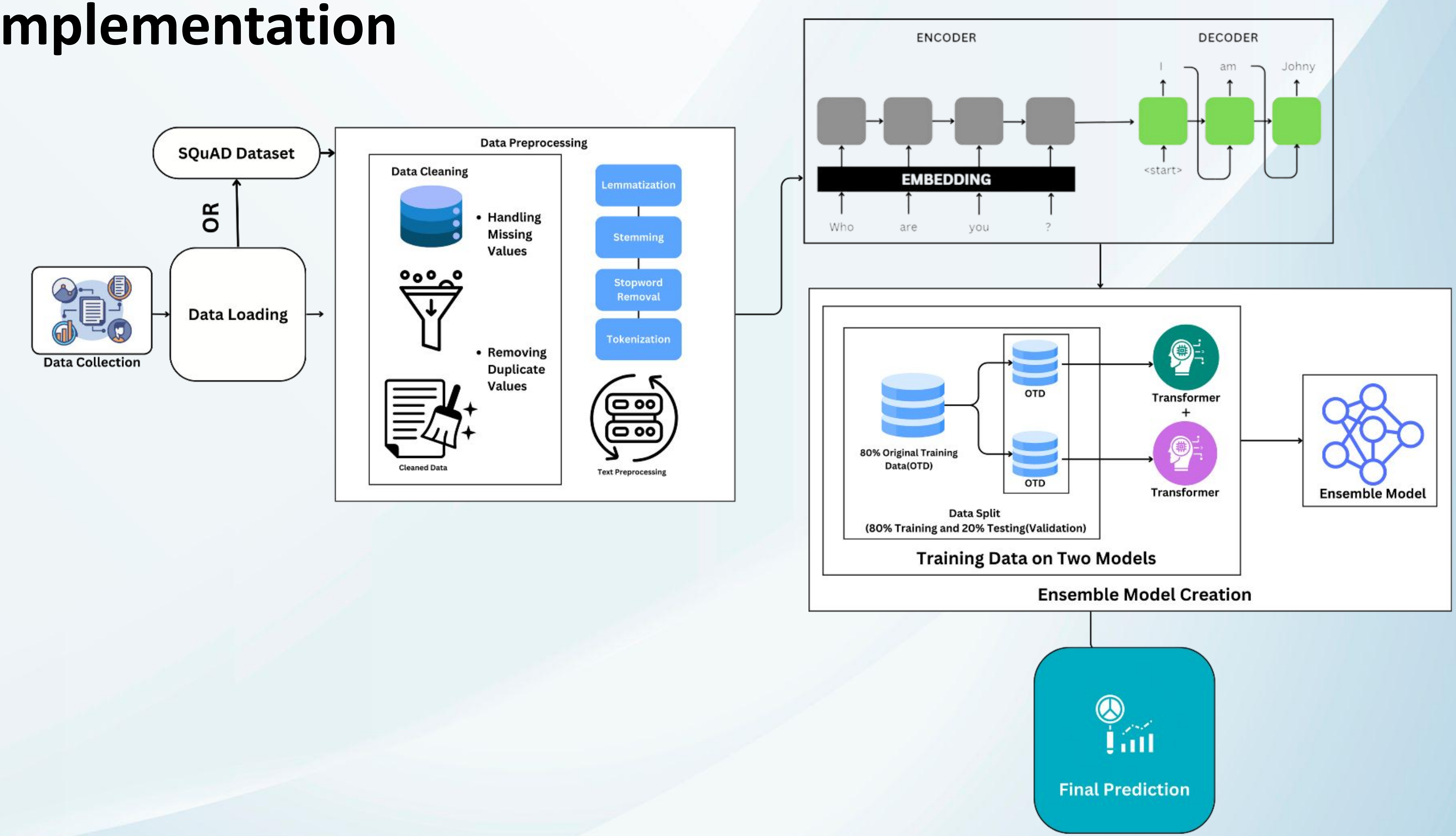
4.Fine-Tuning

- **T5 Small:** Configured inputs as "context: question"; trained for 3 epochs with checkpoints at each epoch.
- **BART Base:** Limited labels to 128 tokens, trained in FP16 mode for efficiency, saved checkpoints each epoch.

5.Ensemble Learning

- Fine-tuned multiple BART base models to create an ensemble.
- Averaged predictions from each model during inference to improve accuracy and reduce variance.

Implementation



Data Collection

Data Loading

OR

SQuAD Dataset

Data Preprocessing

Data Cleaning

- Handling Missing Values
- Removing Duplicate Values

Cleaned Data

Lemmatization

Stemming

Stopword Removal

Tokenization

Text Preprocessing

ENCODER

DECODER

EMBEDDING

Training Data on Two Models

80% Original Training Data(OTD)

OTD

OTD

Data Split
(80% Training and 20% Testing(Validation)

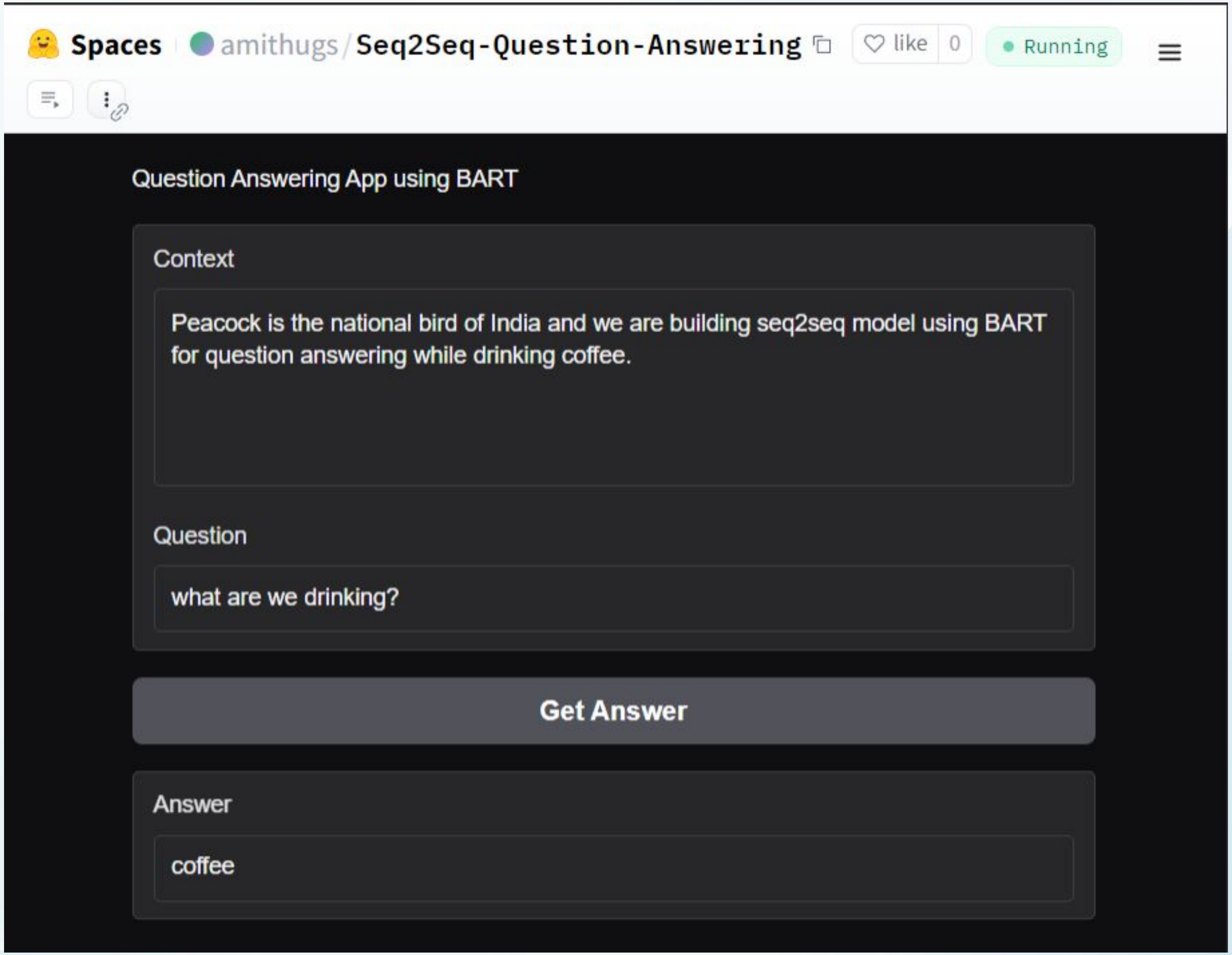
Transformer
+
Transformer

Ensemble Model

Ensemble Model Creation

Final Prediction

OUTPUT



Question Answering App using BART

Context

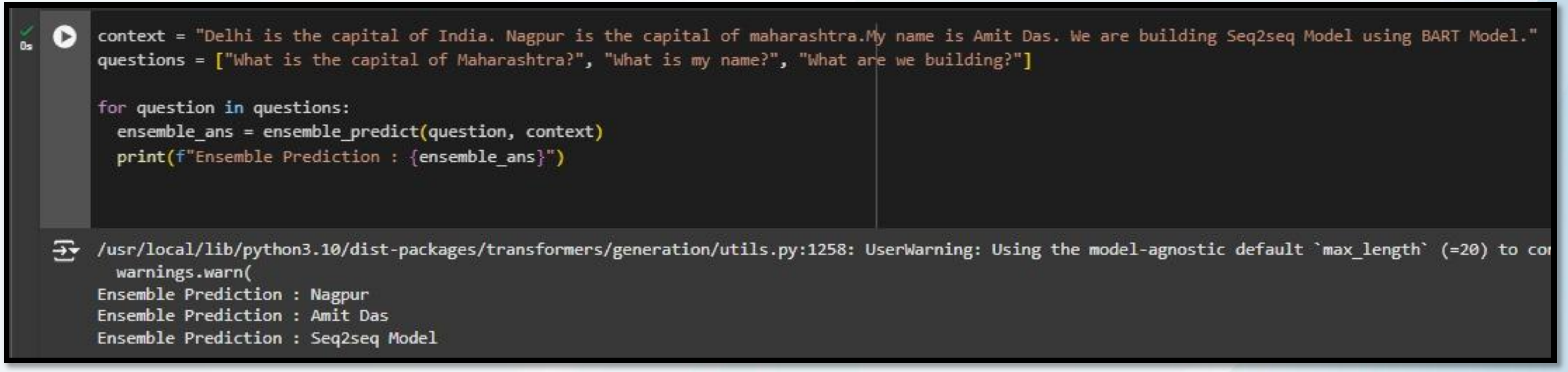
Peacock is the national bird of India and we are building seq2seq model using BART for question answering while drinking coffee.

Question

what are we drinking?

Answer

coffee



```
context = "Delhi is the capital of India. Nagpur is the capital of maharashtra.My name is Amit Das. We are building Seq2seq Model using BART Model."
questions = ["What is the capital of Maharashtra?", "What is my name?", "What are we building?"]

for question in questions:
  ensemble_ans = ensemble_predict(question, context)
  print(f"Ensemble Prediction : {ensemble_ans}")
```

```
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1258: UserWarning: Using the model-agnostic default `max_length` (=20) to cor
  warnings.warn(
Ensemble Prediction : Nagpur
Ensemble Prediction : Amit Das
Ensemble Prediction : Seq2seq Model
```

RESULT

Performance on SQuAD: BART Base outperformed T5 Small in accuracy. BART Ensemble achieved the highest accuracy by averaging predictions, enhancing robustness.

CONCLUSION

Successfully fine-tuned T5 Small and BART Base for question answering. BART Ensemble provided the best results in terms of accuracy and reliability. Key factors for success included effective preprocessing, hyperparameter tuning, and ensemble learning. Transformer models demonstrated strong effectiveness for question answering, with ensemble methods significantly boosting performance.

References

1.Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (Vol. 1, p. 2).

2. Colin, R. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*, 140-1.

3. Clark, K. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

4. Lan, Z. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

5. Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

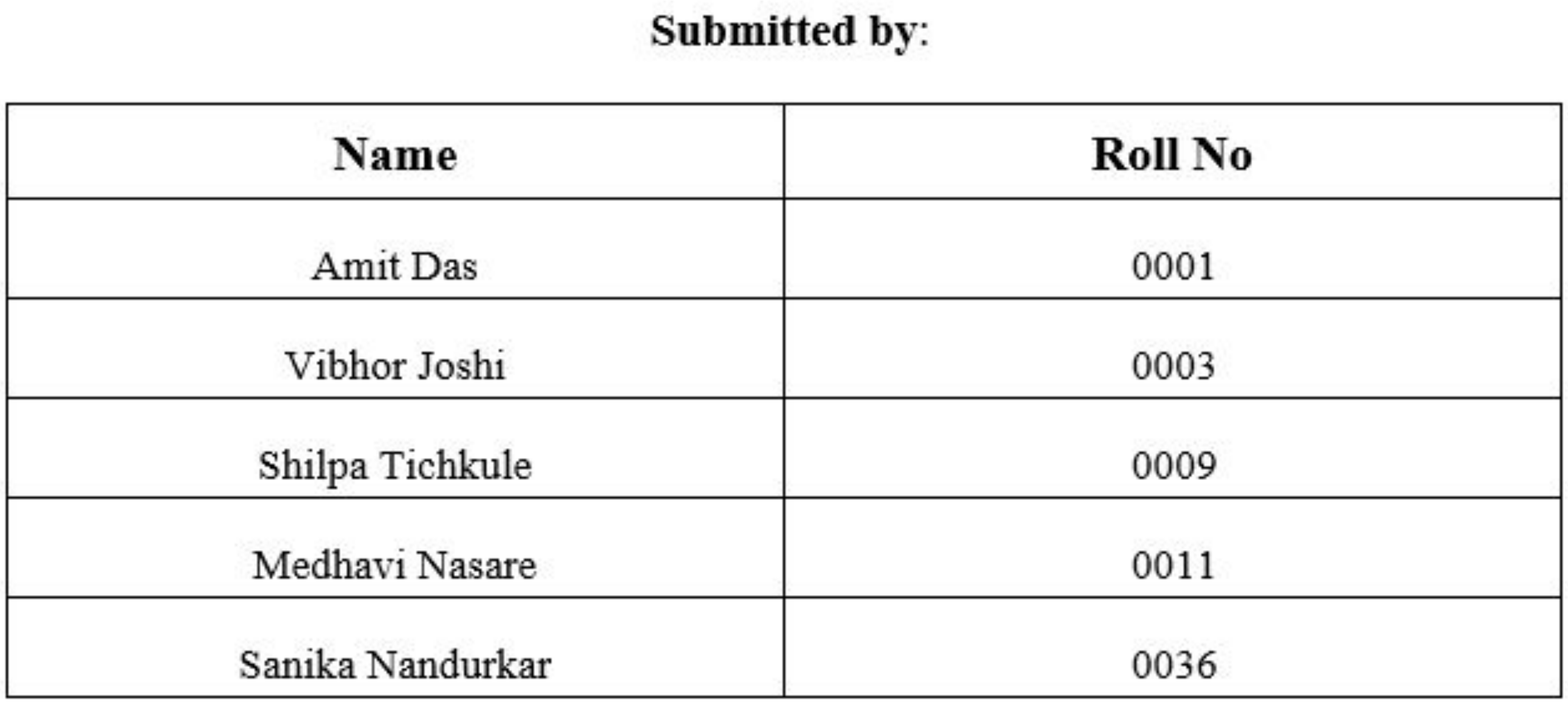
6. Zope, B., Mishra, S., Shaw, K., Vora, D. R., Kotecha, K., & Bidwe, R. V. (2022). Question answer system: A state-of-art representation of quantitative and qualitative analysis. *Big Data and Cognitive Computing*, *6*(4), 109.

7. Abbasiantaeb, Z., & Momtazi, S. (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(6), e1412.

8. Mohammed, S., Shi, P., & Lin, J. (2017). Strong baselines for simple question answering over knowledge graphs with and without neural networks. *arXiv preprint arXiv:1712.01969*.

9.Pandya, H. A., & Bhatt, B. S. (2021). Question answering survey: Directions, challenges, datasets, evaluation matrices. *arXiv preprint arXiv:2112.03572*.

10.Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... & Schulman, J. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.



Submitted by:

Name	Roll No
Amit Das	0001
Vibhor Joshi	0003
Shilpa Tichkule	0009
Medhavi Nasare	0011
Sanika Nandurkar	0036

Under the guidance of

Amit Pandey

G. H. RAISONI COLLEGE OF ENGINEERING,NAGPUR