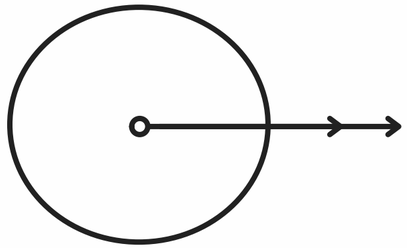
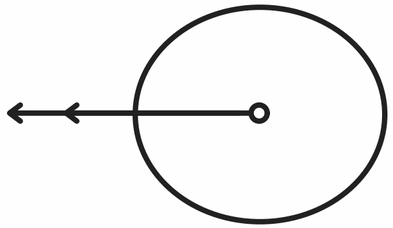


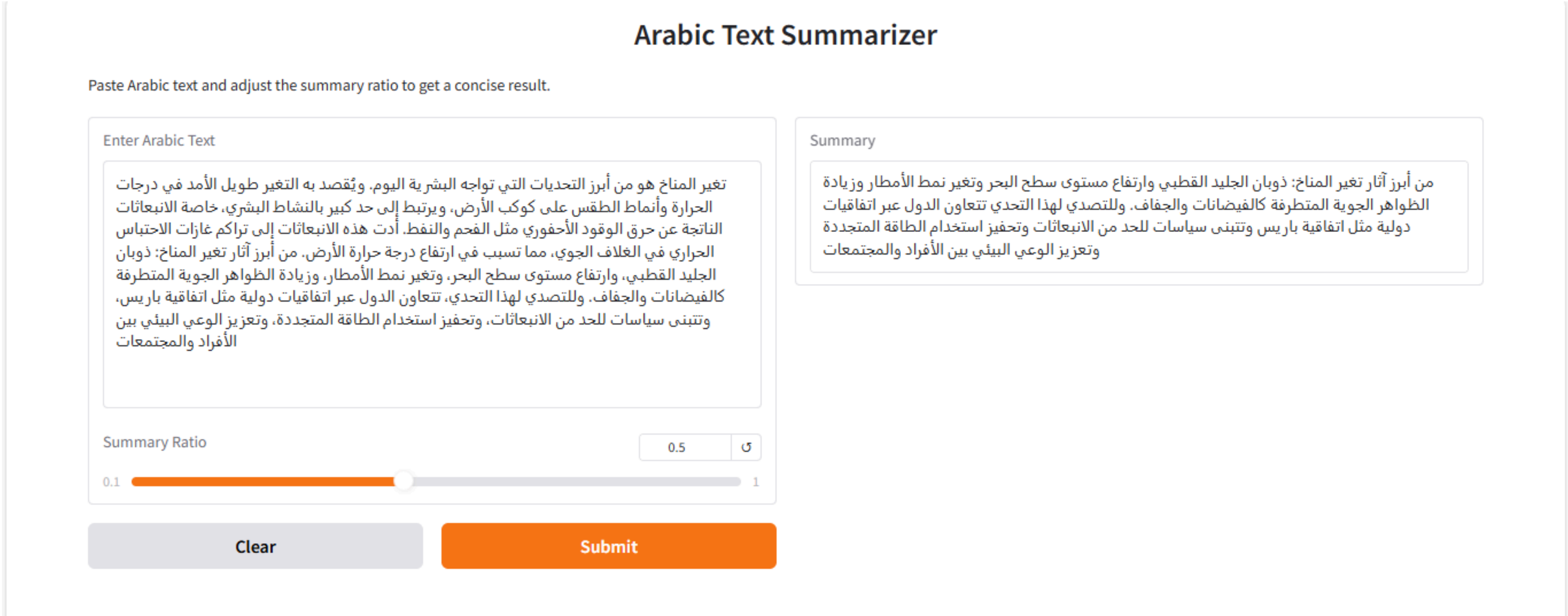
Arabic Text Summarization



01 Project Overview

Project Title: Arabic Text Summarization using AraBART

Goal: Build a model that can generate concise summaries from Arabic text inputs.



02 Dataset

Dataset link: https://www.kaggle.com/datasets/abdalrahmanshahrour/arabicsummarization

Key Features:

- Fields per example:
  - text: The original raw paragraph in Arabic.
  - type: A category or label describing the text type or source (e.g., article, story, etc.). **“not used”**
  - Processed Text: A cleaned and normalized version of the original text, prepared for input into the model (e.g., punctuation cleaned, spacing fixed). **“not used”**
  - summarizer: The human-written reference summary that the model is trained to predict.
- Content Type: The paragraphs include various topics, including cultural, historical, and descriptive content, suitable for training on narrative and informative texts.

03 preprocessing

Text cleaning included:

Removal of diacritics, Latin characters, punctuation.

Normalization of Arabic letters (e.g., ه → ۀ ,ا → ٱ).

Whitespace normalization.

Final columns:

- **Processed Text:** Cleaned text column.
- **summarizer:** Cleaned summary.

04 Model Architecture

- **Base Model:** moussaKam/AraBART (Pretrained on Arabic text)
- **Model Type:** Encoder-Decoder (Seq2Seq)
- **Tokenizer:** AutoTokenizer from AraBART
- **Max Input Length:** 512 tokens
- **Max Output Length:** 110 tokens

05 Training Details

- **Optimizer:** AdamW
- **Loss Function:** CrossEntropyLoss
- **Epochs:** 10
- **Learning Rate:** 5e-5

06 Model Evaluation

- **Metrics Used:**
- **ROUGE-1 / ROUGE-2 / ROUGE-L:** Measures overlap of n-grams between generated and reference summaries.
- **BLEU:** Measures precision of n-grams with brevity penalty.

Metric	Fine-tuned AraBART
ROUGE-1	0.65
ROUGE-2	0.6
ROUGE-L	0.6
BLEU	0.5

07 Model limitation

- Like most transformer-based models, AraBART has a maximum token limit.
- Fine-tuning and inference with AraBART can be resource-intensive (especially on long texts).
- Requires GPU for practical use.