# Convo AI: Natural language processing (UCS664)

# Project on :

Punjabi language Text Summarization

# Submitted By:

|  |  |
| --- | --- |
| Pragya Gupta | 102103407  3CO15 |
| Yash Dogra | 102166002  3CS11 |

**B.E. Third Year – COE**

# Submitted To:

**Dr. Jasmeet Singh**



# Computer Science and Engineering Department Thapar Institute of Engineering and Technology Patiala – 147001

**Dataset: XLSum (Punjabi Language)**

We have used **Extractive Summarization** -It does not generate new text but rather extracts and compiles existing content.

* The [csebuetnlp/xlsum](https://huggingface.co/datasets/csebuetnlp/xlsum) dataset is a comprehensive dataset containing multilingual news articles and their summaries. The dataset includes a wide range of languages. For this project, the XLSum dataset for Punjabi is used from the dataset library is used. There is a 80%-10%-10% split but in Punjabi language we have a **total of 10267 sample counts of which 8215 are for training , 1026 for test and 1026 for validation**.
* from datasets import load\_dataset

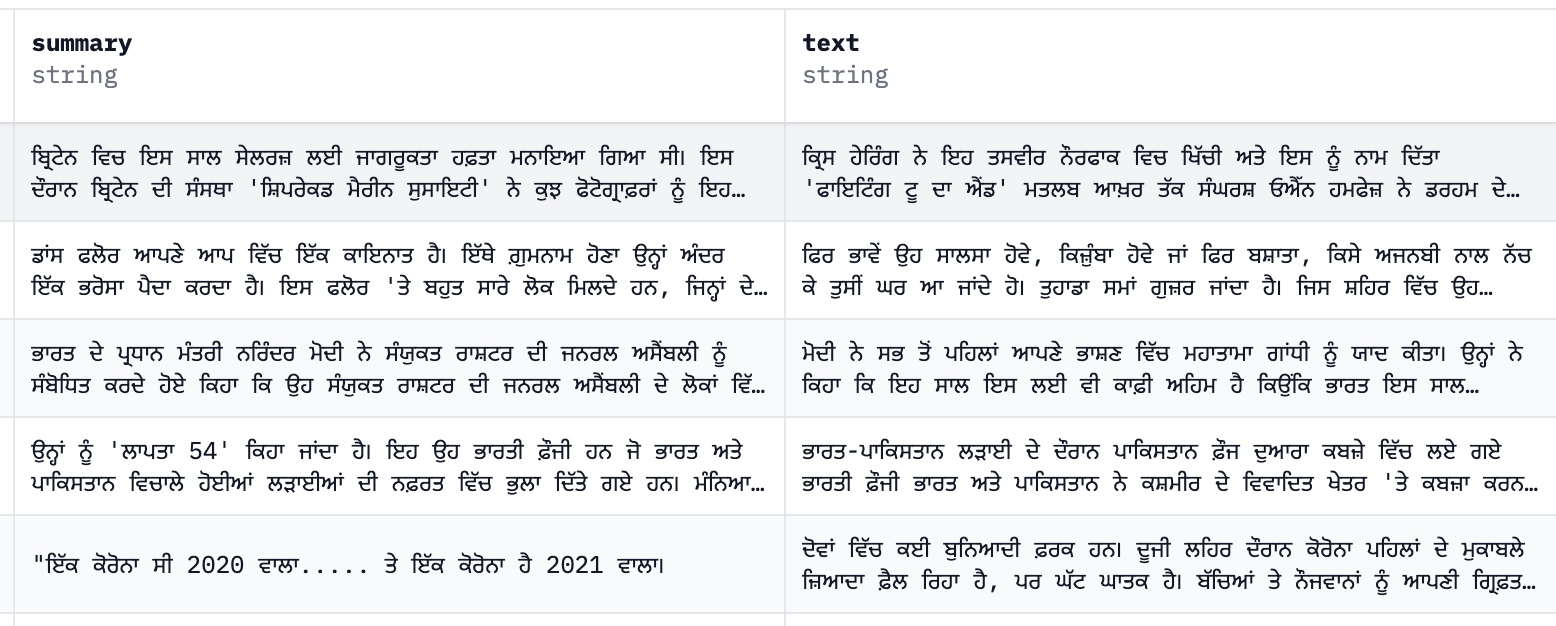
config\_name='punjabi'

dataset = load\_dataset("csebuetnlp/xlsum",config\_name)

**Data Fields**

* 'id': A string representing the article ID.
* 'url': A string representing the article URL.
* 'title': A string containing the article title.
* 'summary': A string containing the article summary.
* 'text' : A string containing the article text.





**Transformer Model: T5**

**(Text-To-Text Transfer Transformer)**

In this project, we utilized Text-To-Text Transfer Transformer (T5) is a pre-trained encoder-decoder model handling all NLP tasks as a unified text-to-text-format where the input and output are always text strings. T5-Small is the checkpoint with 60 million parameters.

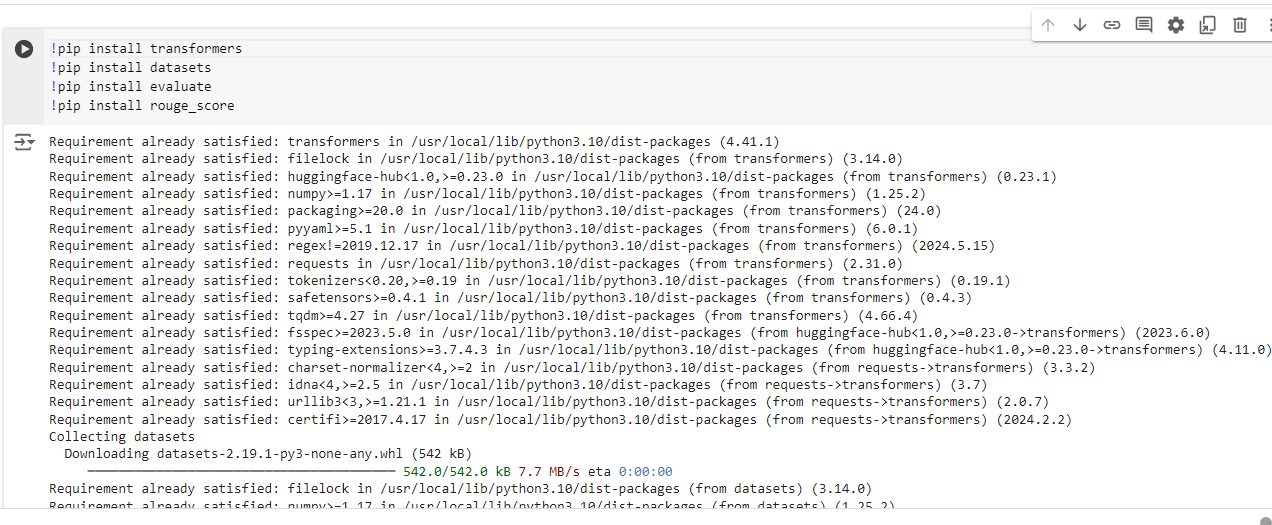
Model Architecture and Features:

* Model Name: T5 (Text-To-Text Transfer Transformer)
* Architecture: Encoder-Decoder architecture, suitable for sequence-to-sequence tasks
* Preprocessing: The input texts are prefixed with "summarize: " before tokenization to indicate the task.
* Optimizer: AdamWeightDecay with a learning rate of 2e-5 and weight decay of 0.01.
* Data Collation: Using DataCollatorForSeq2Seq to dynamically pad inputs and outputs during batching.
* Batch Size: 8 for both training and evaluation.
* Evaluation Metric: **rouge score**- To use it during training, we will create a function compute\_metrics that passes the predictions and labels as a parameter. The length of prediction is added under key ‘gen\_len’

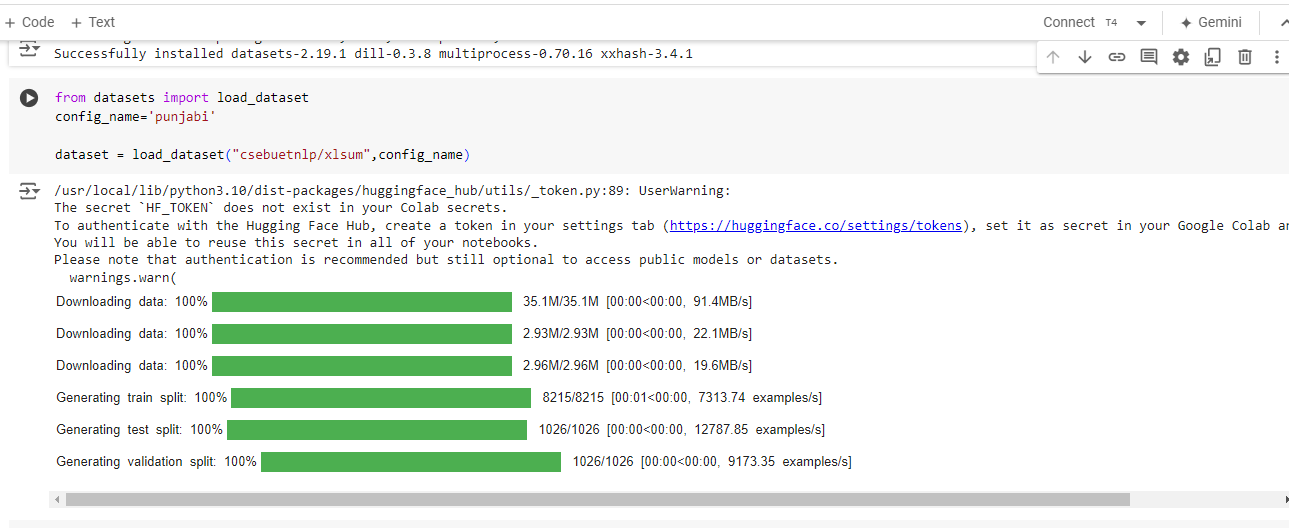
**CODE :**

**Implementation Details:**

Loading Libraries:



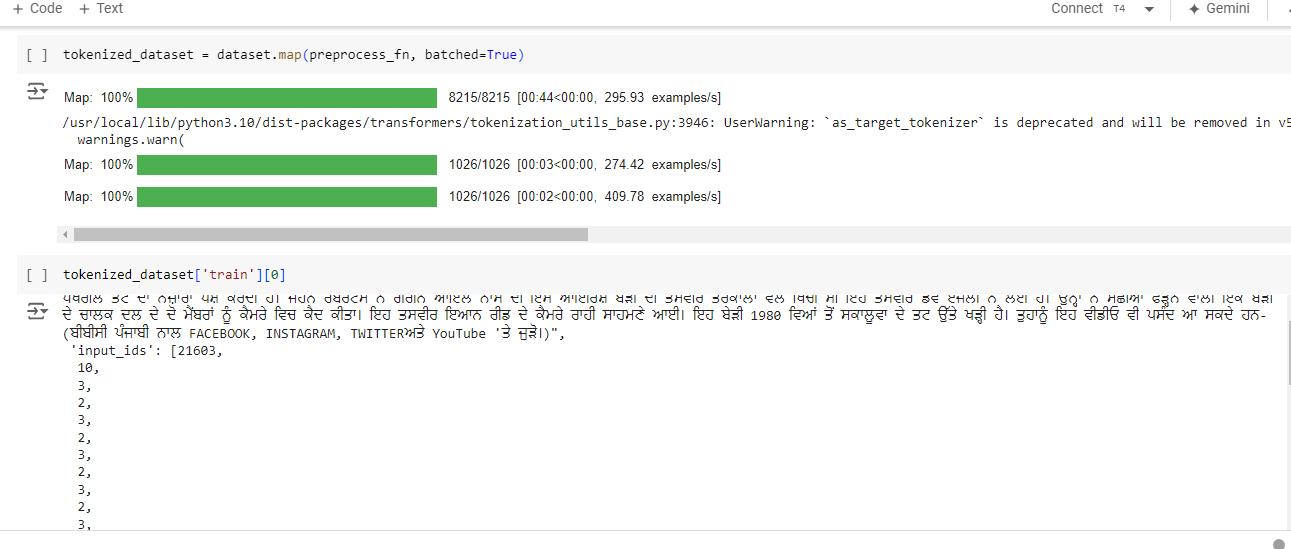
Loading the Dataset:





Using T5-small Transformer model :





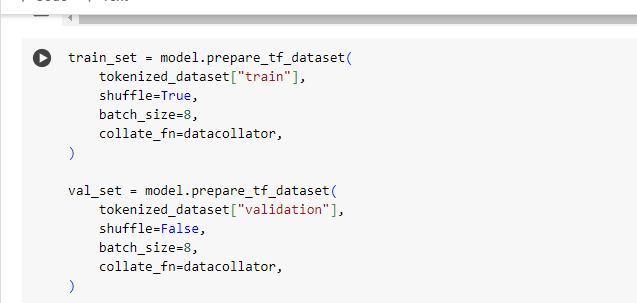
Model and optimizer Initialization:



Breakup of Datasets:

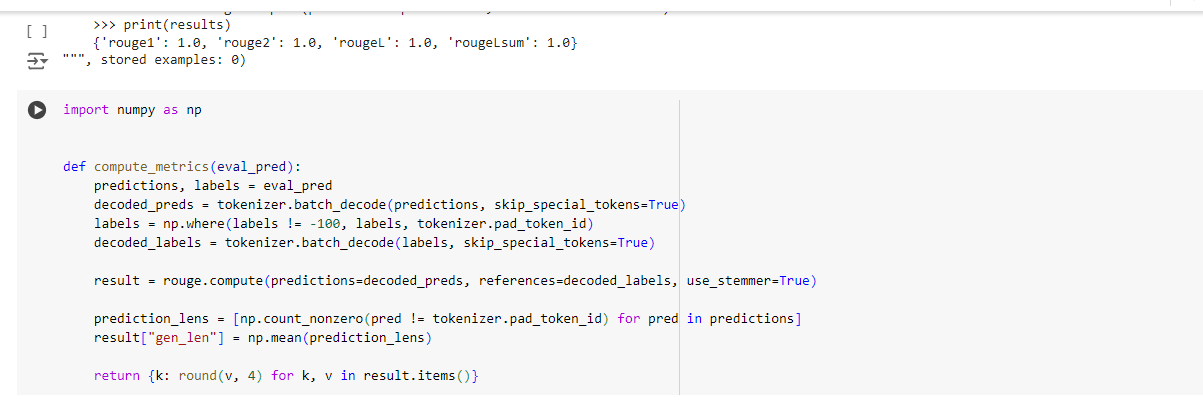
The XLSum dataset, configured for Punjabi, which has training and validation sets to ensure effective training and evaluation of the model.

The validation set is used to tune the model's hyperparameters and evaluate its performance during training. It helps monitor the model's progress and prevents overfitting.

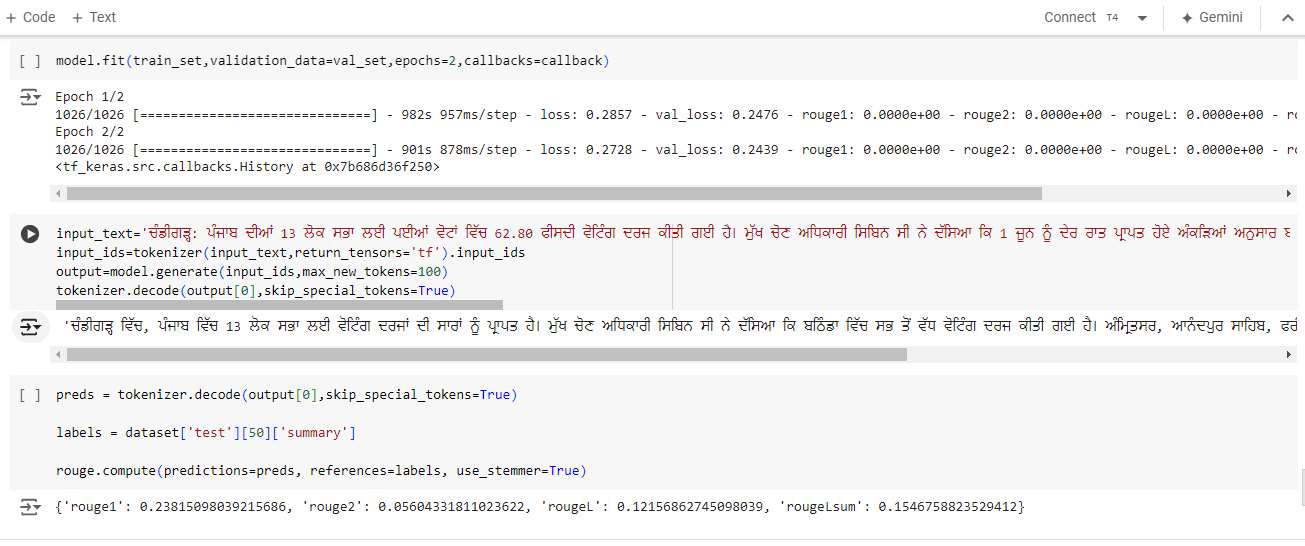


Training:





**Model compilation and Evaluation:**

****

**Results:**

After training the model for 2 epochs, we evaluate its performance using the rouge metric for the 50th entry of test data. The input\_text given is the ‘text’ of dataset[‘test’][50].

The results are as follows:

* {'rouge1': 0.23815098039215686,
* 'rouge2': 0.05604331811023622,
* 'rougeL': 0.12156862745098039,
* 'rougeLsum': 0.1546758823529412}