

# Language Translation System using Neural Networks

MAJOR PROJECT - PPT  
GROUP-157



---

**GUIDE**

Ms. Abhipsa Mahala

---

**DETAILS**

---

**NAME**

**REGN NO.**

---

Saswat Seth

20010352

---

Ashish Kumar Das

20010418

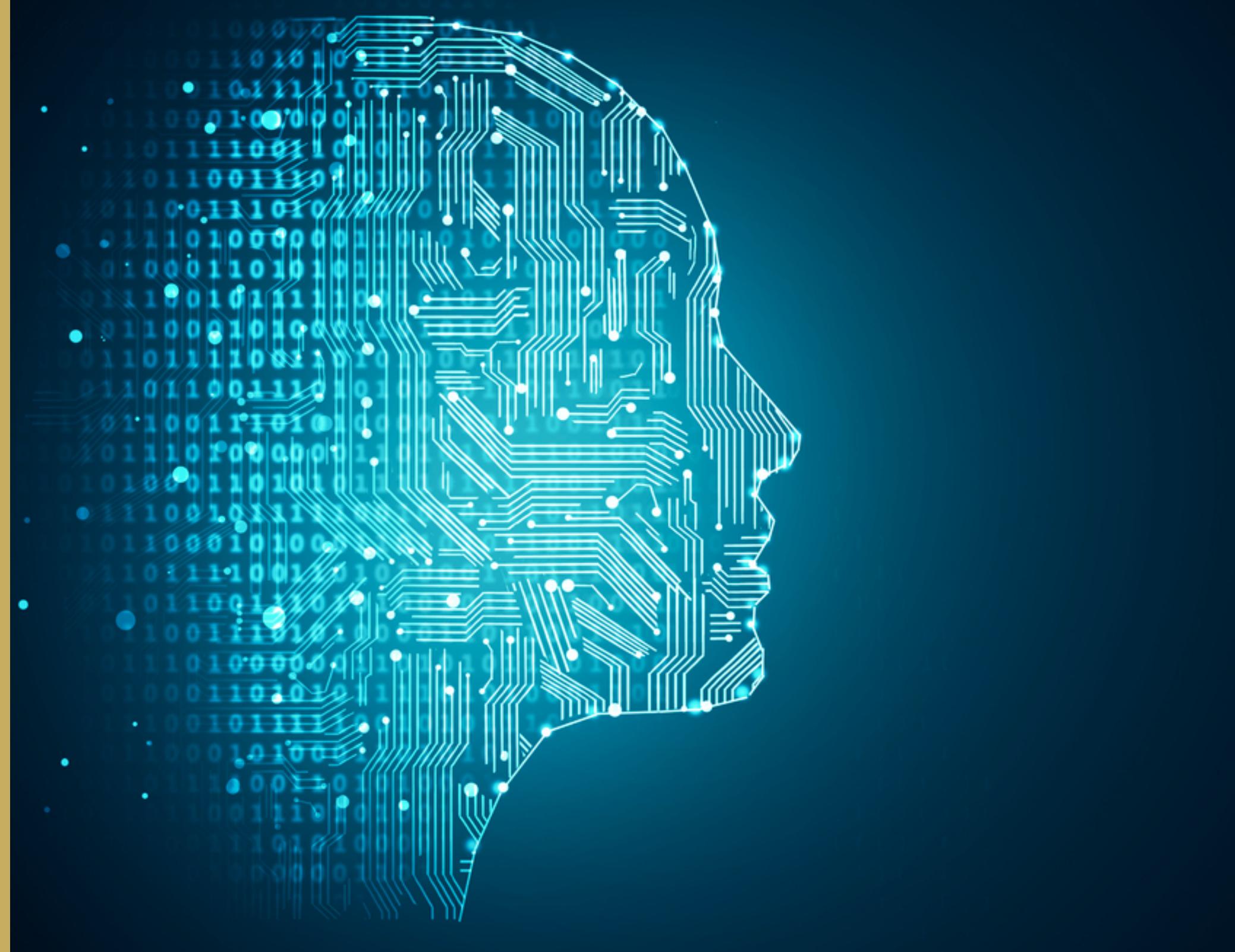
---

SK Sangeet

20010270

# Outlines

<b>Introduction</b>	<b>3-4</b>
<b>Literature Review</b>	<b>5</b>
<b>Methodology</b>	<b>6-11</b>
<b>Results and Discussion</b>	<b>12-13</b>
<b>Appendices</b>	<b>14-17</b>
<b>Summary</b>	<b>18</b>
<b>Conclusion</b>	<b>19</b>
<b>References</b>	<b>20</b>



# Introduction

## Brief Explanation of Neural Machine Translation (NMT):

- NMT is an approach to language translation that utilizes neural networks.
- Unlike traditional rule-based or statistical methods, NMT relies on deep learning to improve translation quality.

## Importance of NMT in Overcoming Language Barriers:

- Breaks down communication barriers by providing accurate and contextually relevant translations.
- Facilitates cross-cultural communication, fostering global collaboration and understanding.



A Language Translation System is a computer-based software or hardware solution that translates text or speech from one language to another. Created through the use of machine learning algorithms, neural networks, and large datasets.

# Introduction

## Swift Evolution of NMT Models:

- NMT has witnessed rapid evolution over the years.
- Continuous advancements in model architectures, training techniques, and pre-training strategies.
- The dynamic landscape reflects the commitment to enhancing translation accuracy and efficiency.
- Highlight the role of advanced models like mT5 in achieving state-of-the-art results.
- Continuous improvements are expanding the capabilities of NMT for diverse language pairs.

## Significance and Relevance

- Multilingual Competence: Facilitates global communication through diverse language translation.
- Fine-tuning Precision: Ensures context-aware accuracy in domain-specific translations.
- Adaptability: Handles varied linguistic nuances, fostering a comprehensive understanding.
- Efficiency: Optimizes performance without compromising quality, thanks to mT5-small's compact design.
- Cross-cultural Communication: Adept at accommodating multiple languages for inclusive communication.
- Context Awareness: Exhibits advanced contextual understanding for improved accuracy.

# Literature Review

Name	Advantages	Disadvantages
• mT5: A massively multilingual pre-trained text-to-text transformer	<ul style="list-style-type: none"><li>• Multilingual capabilities.</li><li>• Versatility in tasks.</li><li>• Improved translation quality.</li></ul>	<ul style="list-style-type: none"><li>• May not excel in language-specific translation.</li><li>• Model size may be a limitation.</li><li>• Complex for specific tasks.</li></ul>
• BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	<ul style="list-style-type: none"><li>• Improved translation quality.</li><li>• Leveraged pre-trained models.</li><li>• Enhanced language understanding.</li></ul>	<ul style="list-style-type: none"><li>• Primarily designed for language understanding.</li><li>• Adaptation may require substantial fine-tuning.</li><li>• Not language-specific.</li></ul>

# Methodology

## How do Transformers work:

Introduction to Transformers: Neural networks from "Attention is All You Need," emphasizing self-attention mechanisms.

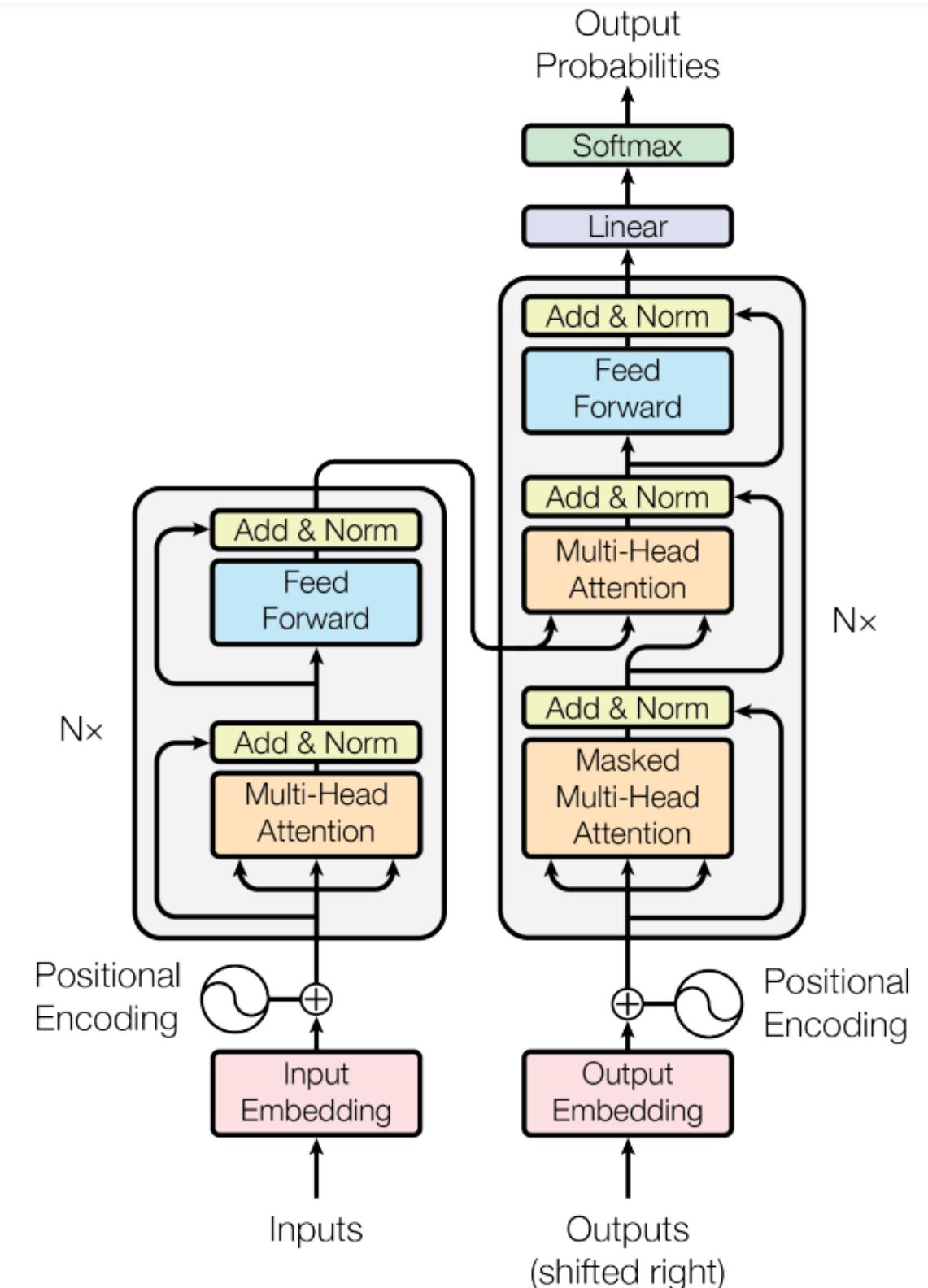
Self-Attention Mechanism: Assigns variable importance to input sequence parts via attention scores.

Multi-Head Attention: Utilizes multiple attention heads concurrently, enhancing overall understanding.

Positional Encoding: Provides crucial information about word positions for effective sequential data handling.

Feedforward Neural Network: Adds depth, capturing non-linear relationships following each attention head.

Layer Normalization and Residual Connections: Incorporates stability elements, easing deep network training.



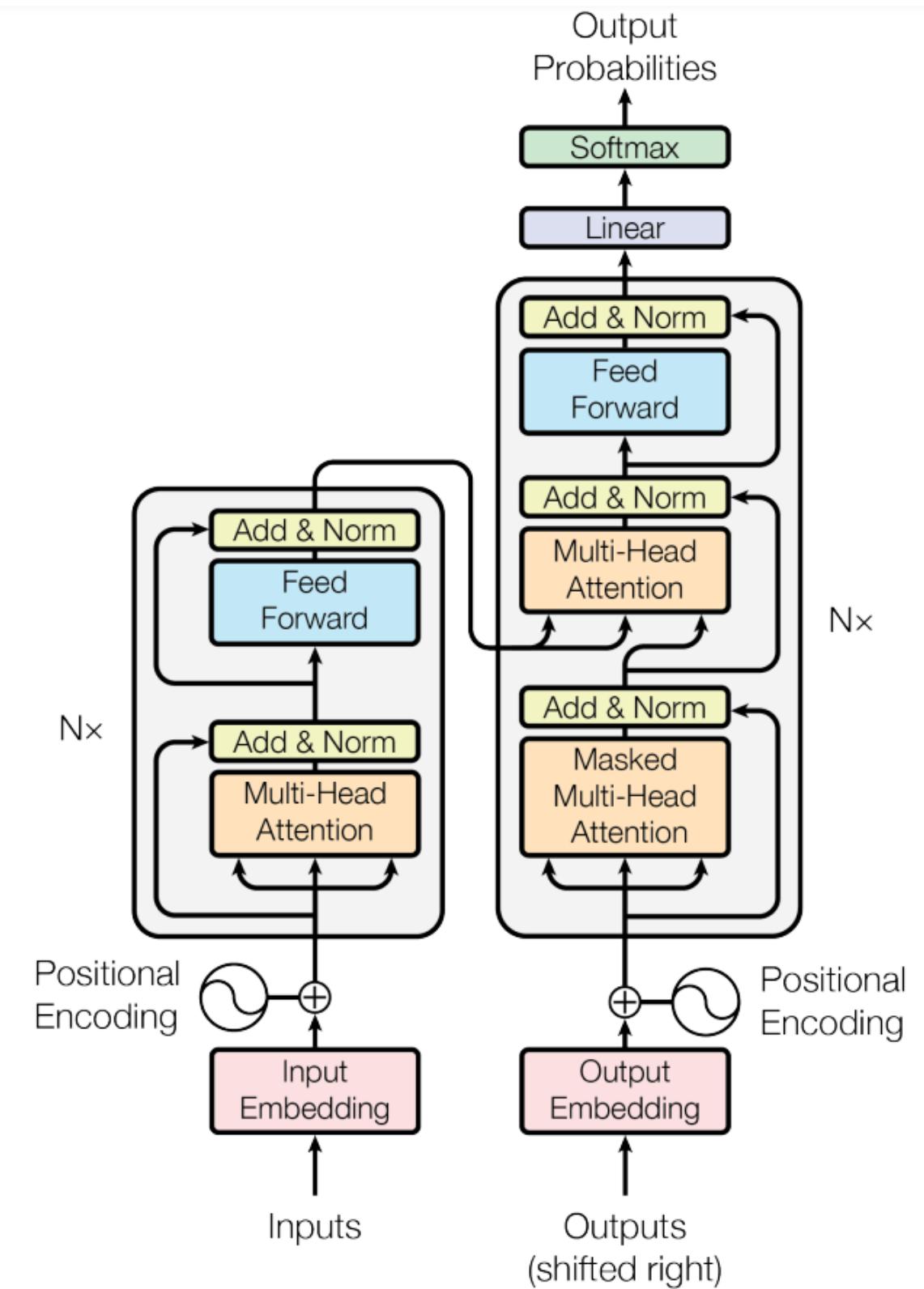
# Methodology

## How do Transformers work(contd.):

Encoder and Decoder: A comprehensive model with an encoder for input processing and a decoder for output generation.

Training and Fine-Tuning: Involves pretraining on extensive datasets and fine-tuning for specific tasks, showcasing adaptability.

Applications: Transformers serve as the backbone for various NLP applications, playing pivotal roles in machine translation, text summarization, and question answering.



# Methodology

## How do Transformers work(contd.):

Input Text: Represents the source text for translation, initiating the process.

Tokenization: Breaks input text into tokens for model understanding.

Embedding Layer: Converts tokens into numerical vectors for model processing

Encoder (Transformers): Encodes input text using transformer architecture.

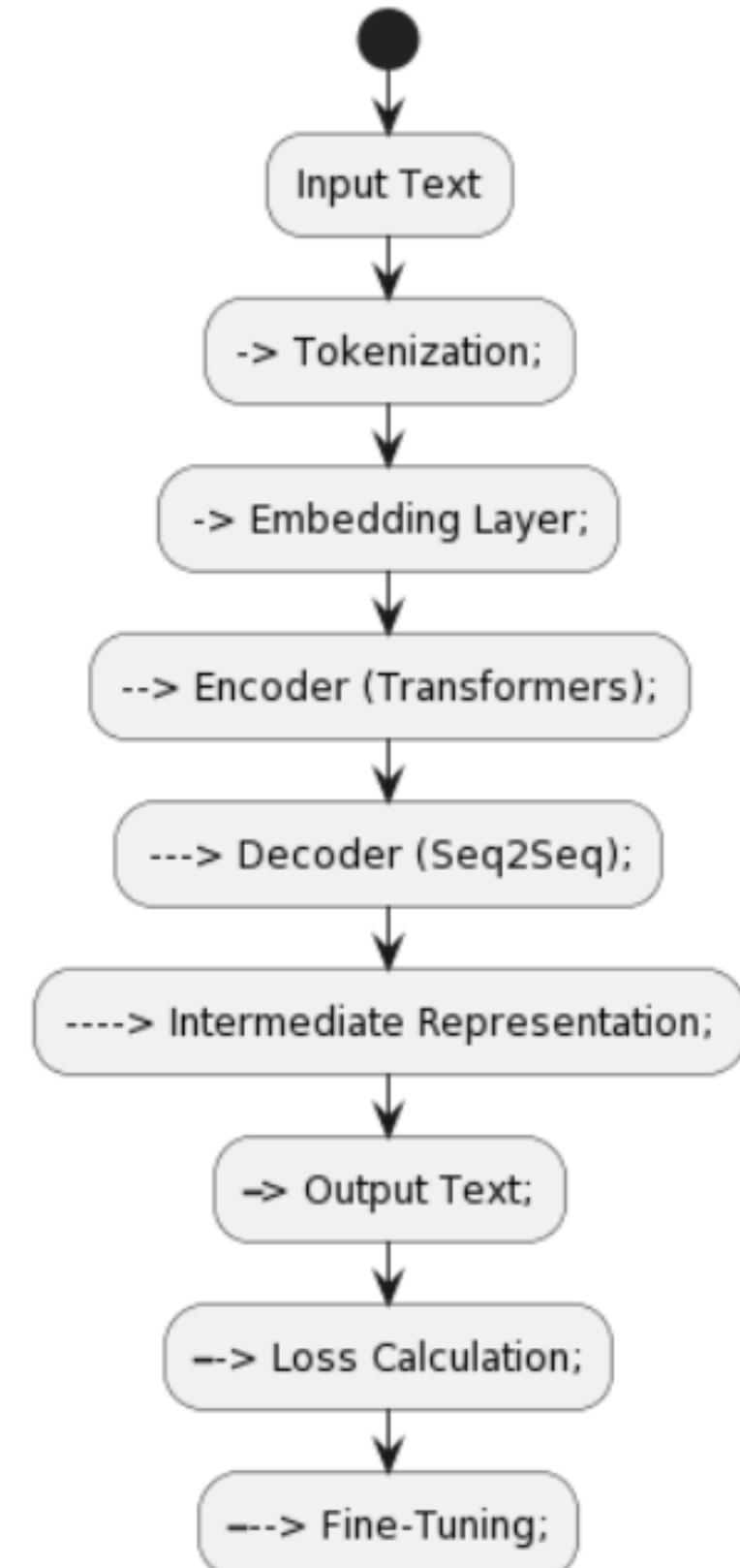
Decoder (Seq2Seq): Generates translated text with Seq2Seq architecture.

Intermediate Representation: Models internal state for output generation.

Output Text: Presents the translated text, the final result.

Loss Calculation: Measures translation accuracy for model optimization.

Fine-Tuning: Adjusts parameters for improved translation performance.



# Methodology

## **Research Design:**

- **Efficient Model Choice:**
  - Selection of 'mt5-small' due to its alignment with project goals and efficiency in multilingual translation tasks.
- **Multilingual Focus for Diversity:**
  - Project emphasis on English, Japanese, and Chinese ensures comprehensive language coverage.
- **Dataset:**
  - 'alt' dataset from Hugging Face chosen for its rich multilingual content, serving as a robust foundation for model training.
- **Supervised Training with Transfer Learning:**
  - Utilization of a supervised training approach with labeled translations.
  - Application of transfer learning techniques to leverage pre-existing knowledge.
- **Ethical Considerations:**
  - Priority on ethical considerations throughout the project, including data privacy.
  - Measures in place to address and mitigate biases in translation outcomes.

# Methodology

## Data Collection Methods:

- Utilization of the 'alt' dataset from the Hugging Face repository, providing diverse multilingual translations for training and evaluation.
- Leveraging the Hugging Face Datasets library for seamless and efficient access to the 'alt' dataset, eliminating the need for manual web scraping.

## Analysis Techniques:

### Tokenization and Text Processing:

- Implementing SentencePiece for effective tokenization and text processing.

### Model Fine-Tuning:

- Fine-tuning 'mt5-small' for precise alignment with translation objectives.

### Evaluation and Comparison:

- Thoroughly assessing model performance and comparing results with existing literature.
- Benchmarking against other translation models for context and relevance.

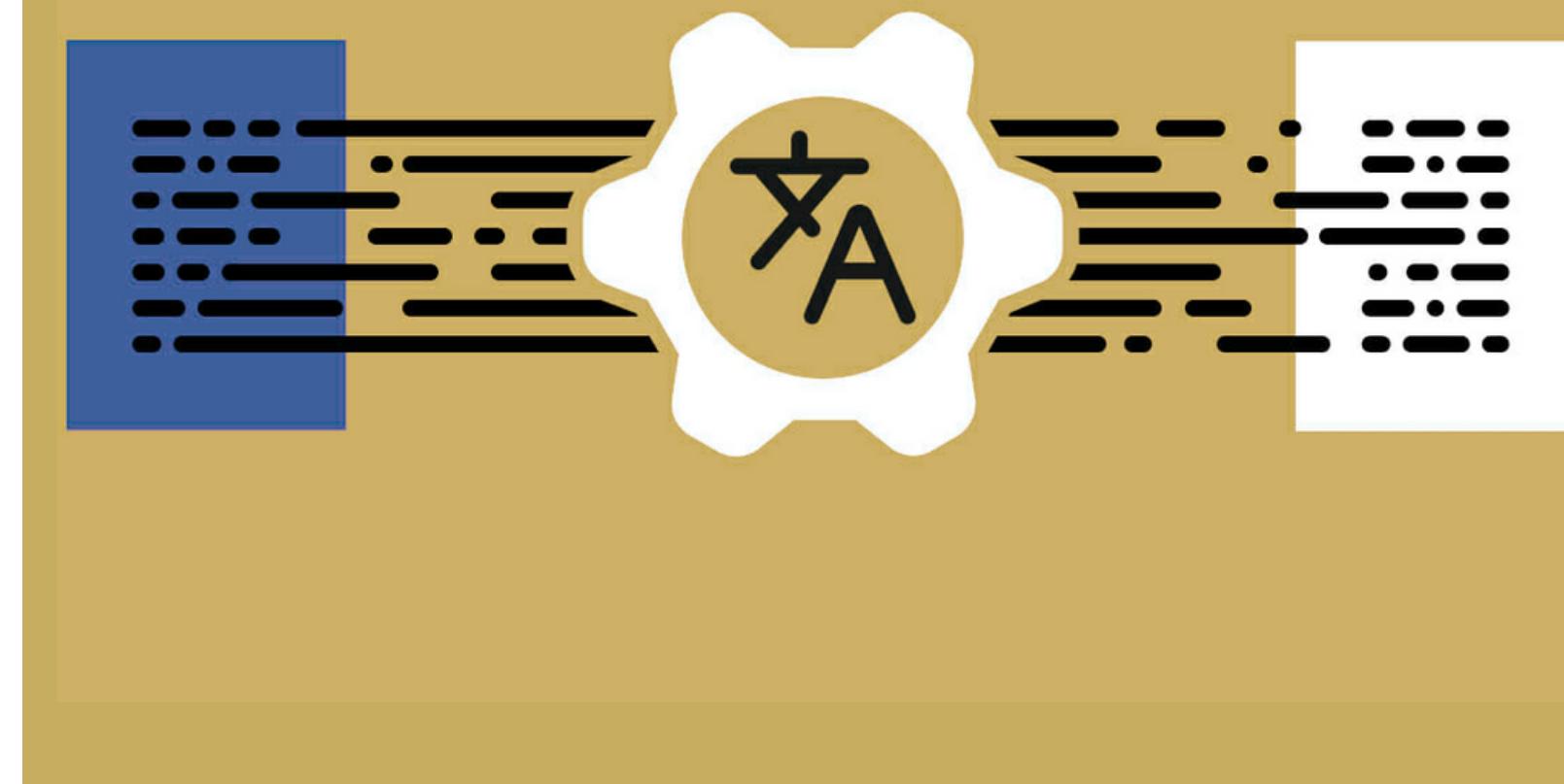
### Exploratory Data Analysis (EDA):

- Preliminary exploration of 'alt' dataset for insights into data distribution and patterns.

# Methodology

## Methodology Limitations:

- Challenges in handling idiomatic expressions and cultural nuances may impact the model's proficiency in capturing subtle linguistic variations.
- The 'alt' dataset's quality and representativeness pose potential biases or gaps, affecting the model's generalization to real-world translation scenarios.
- Sensitivity to input length and complexity could affect performance on longer or more intricate sentences.
- The resource-intensive nature of training and fine-tuning the 'mt5-small' model demands significant computational power and time.
- Reliance on pre-existing translations in the 'alt' dataset may introduce biases or inaccuracies, influencing the quality of generated translations.



# Results and Discussion

NMT.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

RAM Disk

+ Code + Text

[23] Epoch: 3 | Step: 50 | Avg. loss: 2.849 | lr: 0.00016086309523809526  
Epoch: 3 | Step: 100 | Avg. loss: 2.850 | lr: 0.00015342261904761906  
Epoch: 3 | Step: 150 | Avg. loss: 2.812 | lr: 0.00014598214285714288  
Epoch: 3 | Step: 200 | Avg. loss: 2.828 | lr: 0.00013854166666666667  
Epoch: 3 | Step: 250 | Avg. loss: 2.836 | lr: 0.0001311011904761905  
Epoch: 3 | Step: 300 | Avg. loss: 2.793 | lr: 0.00012366871428571428  
Epoch: 3 | Step: 350 | Avg. loss: 2.839 | lr: 0.0001162282380952381  
Epoch: 3 | Step: 400 | Avg. loss: 2.900 | lr: 0.0001087797619047619  
Epoch: 3 | Step: 450 | Avg. loss: 2.855 | lr: 0.00010133928571428571  
Epoch: 3 | Step: 500 | Avg. loss: 2.812 | lr: 9.389880952380952e-05  
Epoch: 3 | Step: 550 | Avg. loss: 2.797 | lr: 8.645833333333334e-05  
Epoch: 3 | Step: 600 | Avg. loss: 2.855 | lr: 7.901785714285714e-05  
Epoch: 3 | Step: 650 | Avg. loss: 2.794 | lr: 7.157738095238095e-05  
Epoch: 3 | Step: 700 | Avg. loss: 2.799 | lr: 6.413690476190476e-05  
Epoch: 3 | Step: 750 | Avg. loss: 2.800 | lr: 5.669642857142857e-05  
Epoch: 3 | Step: 800 | Avg. loss: 2.818 | lr: 4.925595238095238e-05  
Epoch: 3 | Step: 850 | Avg. loss: 2.771 | lr: 4.181547619047619e-05  
Epoch: 3 | Step: 900 | Avg. loss: 2.775 | lr: 3.4375e-05  
Epoch: 3 | Step: 950 | Avg. loss: 2.796 | lr: 2.693452380952381e-05  
Epoch: 3 | Step: 1000 | Avg. loss: 2.798 | lr: 1.949484761904762e-05  
Epoch: 3 | Step: 1050 | Avg. loss: 2.748 | lr: 1.2853571428571429e-05  
Epoch: 3 | Step: 1100 | Avg. loss: 2.790 | lr: 4.6130952380952385e-06  
Test loss of 3.135

[24] # Graph the loss

```
window_size = 50
smoothed_losses = []
for i in range(len(losses)-window_size):
    smoothed_losses.append(np.mean(losses[i:i+window_size]))

plt.plot(smoothed_losses[100:])
```

[`matplotlib.lines.Line2D at 0x7da0414ae860`]

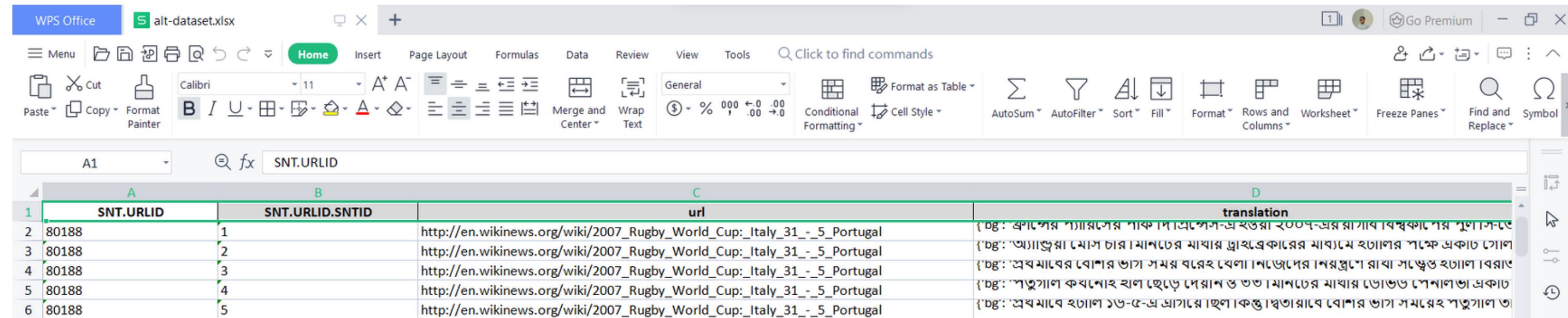
4s completed at 9:39 PM.

# Results and Discussion

- Training Progress: Illustration of the training progress with a decreasing loss graph, reflecting the model's consistent learning and improvement.
- Utilization of 'mt5-small': Highlighting the effective use of the 'mt5-small' model as a powerful neural machine translation architecture, demonstrating commendable performance.
- Epoch-wise Analysis: Examination of different epochs' impact on the model's translation capabilities, providing insights into its evolving understanding.
- Promising Results with Limited Epochs: Discussion on achieving promising translation results with a limited number of epochs, emphasizing the training process's efficiency and the model's quick adaptability.
- Graphical Representation: Inclusion of graphical representations showcasing the loss trend over epochs for a visual understanding of the model's convergence and performance enhancement.a
- Optimizations and Future Directions: Discussion on potential optimizations for further improving the model's performance and exploring avenues for future enhancements and research directions.

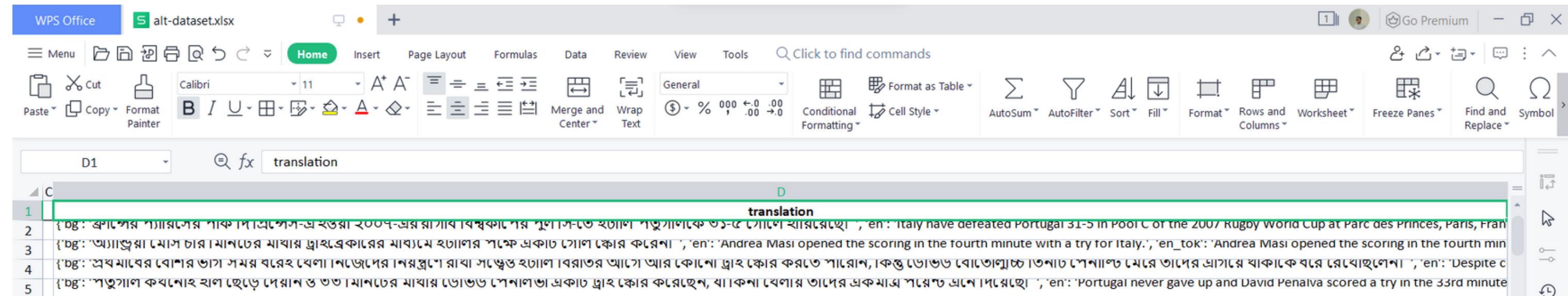
# Appendices

## Dataset:



The screenshot shows a WPS Office spreadsheet window titled "alt-dataset.xlsx". The ribbon menu is visible at the top, and the "Home" tab is selected. The spreadsheet contains four columns: A, B, C, and D. Column A is labeled "SNT.URLID" and column B is labeled "SNT.URLID.SNTID". Column C is labeled "url" and column D is labeled "translation". The data consists of six rows, each containing a URL from the 2007 Rugby World Cup and its corresponding translation in Bulgarian and English.

A	B	C	D
SNT.URLID	SNT.URLID.SNTID	url	translation
80188	1	http://en.wikinews.org/wiki/2007_Rugby_World_Cup:_Italy_31_-_5_Portugal	{'bg': 'България със забележителен старт в група А на световното по регби в Париж. Маси отвори счёта за Италия', 'en': 'Italy have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, Andrea Masi opening the scoring in the fourth minute with a try for Italy.'}
80188	2	http://en.wikinews.org/wiki/2007_Rugby_World_Cup:_Italy_31_-_5_Portugal	{'bg': 'Андреа Маси отвори счёта за Италия с тройка в четвъртия минута', 'en': 'Andrea Masi opened the scoring in the fourth minute with a try for Italy.'}
80188	3	http://en.wikinews.org/wiki/2007_Rugby_World_Cup:_Italy_31_-_5_Portugal	{'bg': 'Италия със забележителен старт в група А на световното по регби в Париж. Маси отвори счёта за Италия', 'en': 'Italy have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, Andrea Masi opening the scoring in the fourth minute with a try for Italy.'}
80188	4	http://en.wikinews.org/wiki/2007_Rugby_World_Cup:_Italy_31_-_5_Portugal	{'bg': 'Португалия със забележителен старт в група А на световното по регби в Париж. Пенаува отвори счёта за Португалия с тройка в 33-тия минута', 'en': 'Portugal have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, David Penalva opening the scoring in the 33rd minute with a try for Portugal.'}
80188	5	http://en.wikinews.org/wiki/2007_Rugby_World_Cup:_Italy_31_-_5_Portugal	{'bg': 'Италия със забележителен старт в група А на световното по регби в Париж. Маси отвори счёта за Италия с тройка в четвъртия минута', 'en': 'Italy have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, Andrea Masi opening the scoring in the fourth minute with a try for Italy.'}



The screenshot shows a WPS Office spreadsheet window titled "alt-dataset.xlsx". The ribbon menu is visible at the top, and the "Home" tab is selected. The spreadsheet contains two columns: C and D. Column C is labeled "translation" and column D is also labeled "translation". The data consists of five rows, each containing a sentence from the 2007 Rugby World Cup and its corresponding translation in Bulgarian and English.

C	D
{'bg': 'България със забележителен старт в група А на световното по регби в Париж. Маси отвори счёта за Италия', 'en': 'Italy have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, Andrea Masi opening the scoring in the fourth minute with a try for Italy.'}	translation
{'bg': 'Андреа Маси отвори счёта за Италия с тройка в четвъртия минута', 'en': 'Andrea Masi opened the scoring in the fourth minute with a try for Italy.'}	translation
{'bg': 'Италия със забележителен старт в група А на световното по регби в Париж. Маси отвори счёта за Италия', 'en': 'Italy have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, Andrea Masi opening the scoring in the fourth minute with a try for Italy.'}	translation
{'bg': 'Португалия със забележителен старт в група А на световното по регби в Париж. Пенаува отвори счёта за Португалия с тройка в 33-тия минута', 'en': 'Portugal have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, David Penalva opening the scoring in the 33rd minute with a try for Portugal.'}	translation
{'bg': 'Италия със забележителен старт в група А на световното по регби в Париж. Маси отвори счёта за Италия с тройка в четвъртия минута', 'en': 'Italy have started Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France with a dominant performance, Andrea Masi opening the scoring in the fourth minute with a try for Italy.'}	translation

# Appendices

## Dataset:

### Data Fields

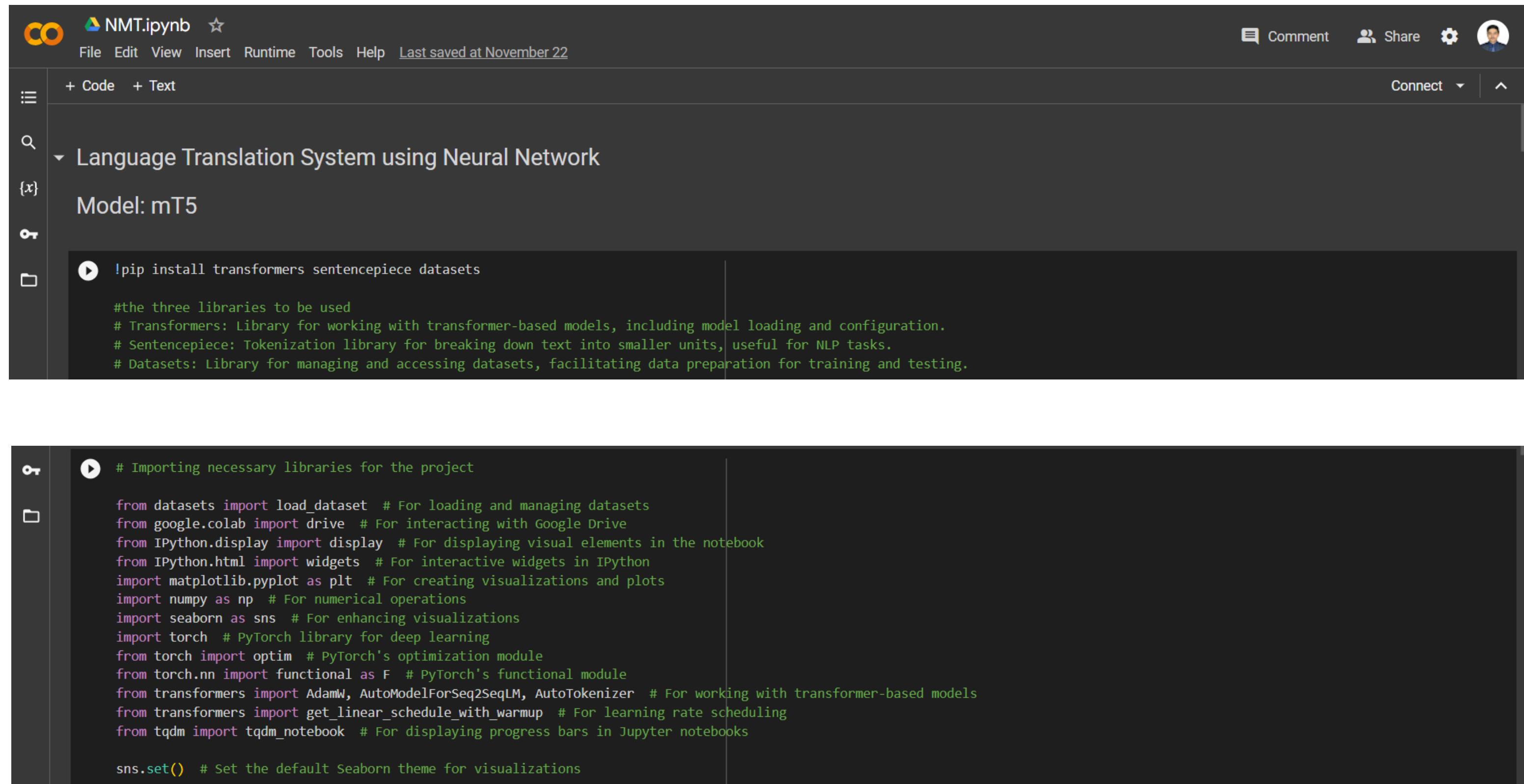
- ALT Parallel Corpus
- SNT.URLID: URL link to the source article listed in URL.txt
- SNT.URLID.SNTID: index number from 1 to 20000. It is a selected sentence from SNT.URLID
- and bg, en, fil, hi, id, ja, khm, lo, ms, my, th, vi, zh correspond to the target language

## Model:

### Mt5-small

- Introduction of T5 (Text-to-Text Transfer Transformer) as a recent advancement in NLP, achieving state-of-the-art results in English-language tasks.
- Introduction of mT5, a multilingual variant of T5, pre-trained on a new Common Crawl-based dataset covering 101 languages.
- Description of the design and modified training process of mT5 to cater to multilingual capabilities.
- Demonstration of mT5's state-of-the-art performance on various multilingual benchmarks, showcasing its versatility and effectiveness.
- Public availability of both the code and model checkpoints used in this work, promoting transparency and accessibility for further research and application.

# Appendices



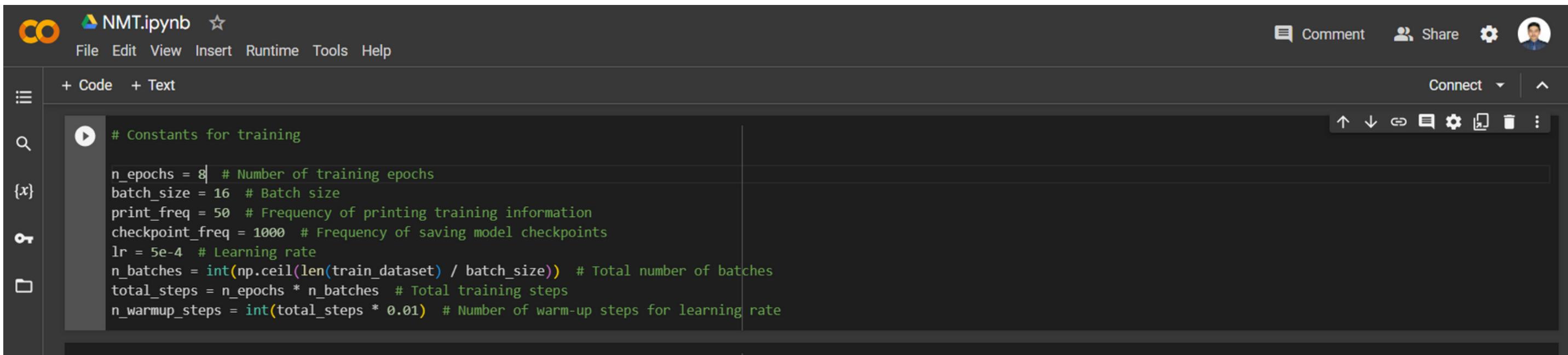
The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** NMT.ipynb
- File Bar:** File Edit View Insert Runtime Tools Help Last saved at November 22
- Toolbar:** Comment Share Settings User Profile
- Sidebar:** + Code + Text, Search, {x}, Model: mT5
- Content Area:**
  - Section:** Language Translation System using Neural Network
  - Text:** Model: mT5
  - Code Cell:** !pip install transformers sentencepiece datasets  
#the three libraries to be used  
# Transformers: Library for working with transformer-based models, including model loading and configuration.  
# Sentencepiece: Tokenization library for breaking down text into smaller units, useful for NLP tasks.  
# Datasets: Library for managing and accessing datasets, facilitating data preparation for training and testing.
- Code Block:** # Importing necessary libraries for the project  

```
from datasets import load_dataset # For loading and managing datasets
from google.colab import drive # For interacting with Google Drive
from IPython.display import display # For displaying visual elements in the notebook
from IPython.html import widgets # For interactive widgets in IPython
import matplotlib.pyplot as plt # For creating visualizations and plots
import numpy as np # For numerical operations
import seaborn as sns # For enhancing visualizations
import torch # PyTorch library for deep learning
from torch import optim # PyTorch's optimization module
from torch.nn import functional as F # PyTorch's functional module
from transformers import AdamW, AutoModelForSeq2SeqLM, AutoTokenizer # For working with transformer-based models
from transformers import get_linear_schedule_with_warmup # For learning rate scheduling
from tqdm import tqdm_notebook # For displaying progress bars in Jupyter notebooks

sns.set() # Set the default Seaborn theme for visualizations
```

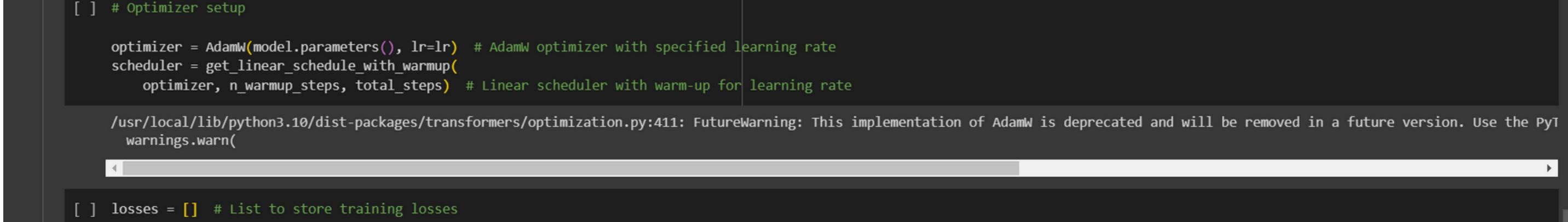
# Appendices



The screenshot shows the first cell of a Jupyter Notebook named 'NMT.ipynb'. The cell contains Python code for defining training constants:

```
# Constants for training

n_epochs = 8 # Number of training epochs
batch_size = 16 # Batch size
print_freq = 50 # Frequency of printing training information
checkpoint_freq = 1000 # Frequency of saving model checkpoints
lr = 5e-4 # Learning rate
n_batches = int(np.ceil(len(train_dataset) / batch_size)) # Total number of batches
total_steps = n_epochs * n_batches # Total training steps
n_warmup_steps = int(total_steps * 0.01) # Number of warm-up steps for learning rate
```



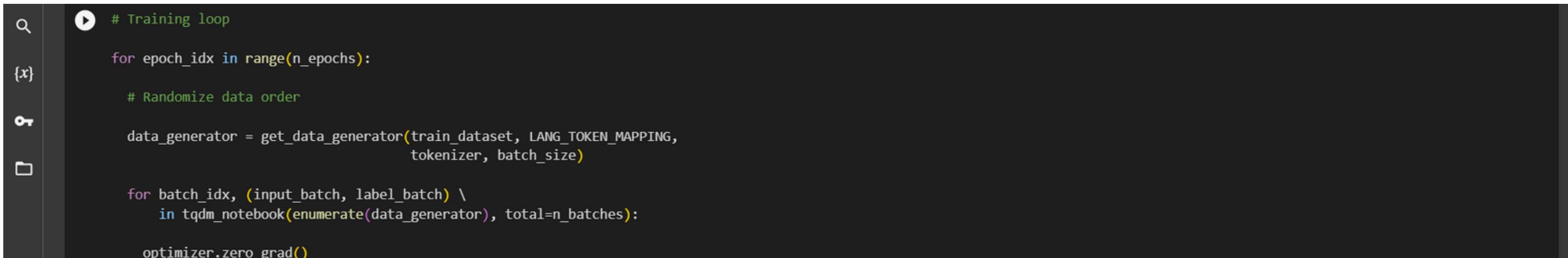
The screenshot shows the second cell of the same Jupyter Notebook. It contains code for optimizer setup and includes a warning message:

```
[ ] # Optimizer setup

optimizer = AdamW(model.parameters(), lr=lr) # AdamW optimizer with specified learning rate
scheduler = get_linear_schedule_with_warmup(
    optimizer, n_warmup_steps, total_steps) # Linear scheduler with warm-up for learning rate

/usr/local/lib/python3.10/dist-packages/transformers/optimization.py:411: FutureWarning: This implementation of AdamW is deprecated and will be removed in a future version. Use the PyT
warnings.warn(
```

The cell ends with a partially visible line: [ ] losses = [] # List to store training losses



The screenshot shows the third cell of the Jupyter Notebook, which contains the main training loop:

```
[ ] # Training loop

for epoch_idx in range(n_epochs):
    # Randomize data order

    data_generator = get_data_generator(train_dataset, LANG_TOKEN_MAPPING,
                                         tokenizer, batch_size)

    for batch_idx, (input_batch, label_batch) \
        in tqdm_notebook(enumerate(data_generator), total=n_batches):
        optimizer.zero_grad()
```

# Summary

- Project Focus: Developing a Language Translation System using Neural Networks.
- Objectives: Enhancing precision and context awareness for specific language pairs.
- Innovation: Convergence of innovation and advancement in cross-lingual communication.
- Approach: Utilizing cutting-edge neural machine translation and sequence-to-sequence models.
- Language Support: Commitment to a wide array of languages and real-time translation capabilities.
- Applications: Enabling transformative applications in diverse industries.
- Impact: Working towards an interconnected and inclusive global society  
Promise: Signifying a linguistic evolution with Neural Networks.

# Conclusion

## Achievements:

- Implemented NMT using mt5 model with alt dataset for language translations.

## Dataset Insights:

- Explored alt dataset, gaining insights into language indicators and translation pairs.

## Model Selection:

- Selected 'mt5-small' model, integrating Transformers, SentencePiece, and Datasets.

## Training Process:

- Employed AdamW optimizer and linear scheduler for iterative model refinement.

## Evaluation and Validation:

- Evaluated model on test dataset, validated through dynamic loss graph.

## Practical Application:

- Developed user-friendly translator tool for input-output translations.

## Future Directions:

- Identified potential for fine-tuning, emphasizing continuous improvement based on user feedback.

# References

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems (pp. 30-40).
- Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., ... Cherry, C. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019.
- Liu, Z., Indra Winata, G., Madotto, A., & Fung, P. (2020). Exploring fine-tuning techniques for pre-trained cross-lingual models via continual learning. arXiv preprint arXiv:2004.14218.
- Carmo, D., Piau, M., Campiotti, I., Nogueira, R., & Lotufo, R. (2020). PTT5: Pretraining and validating the t5 model on Brazilian Portuguese data. arXiv preprint arXiv:2008.09144.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.