1. Pre-processing Module

In the case of T5 model, preprocessing is an essential step to ensure that the input data is in the appropriate format for the model to process effectively. Here's a general outline of the preprocessing steps typically performed before feeding data into the T5 model:

1. **Tokenization**: The input text is tokenized into subword units using a Byte Pair Encoding (BPE) tokenizer or a similar method. This breaks the text into smaller meaningful units, which helps the model to better understand and process the text.
2. **Special Tokens Addition**: Special tokens like the start-of-sentence token, end-of-sentence token, and padding tokens are added to the tokenized input. These tokens help the model understand the beginning and end of the input text and handle variable input lengths.
3. **Text Normalization**: This step involves converting the input text into a standard format by performing operations such as lowercasing, handling punctuation, and removing unnecessary characters. This ensures consistency and reduces noise in the input data.
4. **Padding and Truncation**: The input sequences are either padded with a special padding token to make all sequences of equal length or truncated to a predefined maximum length. This step ensures that the input data is of uniform length, which is necessary for efficient batch processing.
5. **Data Encoding**: The tokenized input text is converted into numerical values that can be processed by the T5 model. This typically involves mapping each token to a unique numerical identifier to create the input data in a format suitable for the model.

2. Definition of simplification (feature-wise)

- Text Simplification: This involves reducing the complexity of the textual data, such as removing complex vocabulary, restructuring sentences for clarity, or paraphrasing complex phrases to make the text more understandable.
- Metadata Reduction: Simplifying the metadata could involve removing redundant or irrelevant metadata fields, or aggregating certain metadata to create more concise and informative features.

3. Scope of simplification (that you would want to address)

- Data Cleaning can be a scope for further improvement. This may include identifying and removing or correcting errors, inconsistencies, or outliers in the dataset, ensuring that the data is of high quality and suitable for analysis or modeling.
- Further, if we have high computation power then we can also use FALN T5, the LLM version of T5.

4. Problem formulation (i.e. precise formulation of Input, Output, Loss function)

Input Format: The T5 model accepts input in the form of a text string, where the input task is provided as a prefix followed by a delimiter (usually a special token like ">>"). The input task can be a wide range of natural language processing tasks such as translation, summarization, question answering, and more.

The T5 (Text-to-Text Transfer Transformer) model has a specific format for both input and output, which contributes to its flexibility in handling various text-based tasks.

Input Format: The T5 model accepts input text along with the input task, where the input task is provided as a prefix.

Output Format: The T5 model generates simplified text as the output.

Loss function: cross-entropy loss

5. Methodology at a broad level (type of model, training methodology)

T5 follows encoder-decoder transformer architecture. It

trained in two stages:

1. **Pretraining**: In the pretraining phase, the T5 model is trained on a diverse and extensive corpus of text data using unsupervised learning. During pretraining, the model learns general language representations and patterns by processing large amounts of text data, which helps it understand the structure and semantics of natural language.
2. **Fine-Tuning**: After pretraining on a large dataset, the T5 model can be fine-tuned on specific downstream tasks, such as translation, summarization, question answering, and more. Fine-tuning involves training the model on task-specific datasets with labeled examples, enabling the model to adapt and specialize its knowledge and capabilities to the particular task at hand. *In our case*, we can further fine-tune the model on the text simplification dataset.

.

6. Timeline for Completion of Project:

Completed:

- Data collection and pre-processing.

Remaining:

6th - 11th Nov

- Lingustic Study

12th - 18th Nov

- Model development and training.
- Model evaluation and fine-tuning.

19th - 22th Nov

- Documentation, testing, and optimization.
- Final testing, deployment, and report generation.

7. Task Delegation Among Members:

Linguistic Research:

All 4 team mates

Data Work:

Mukesh & Shreya

Model Creating & Training

Mukesh & Ganesh

Finetuning

Arjun & Mukesh

Deployment & Documentation

Ganesh & Shreya