End-to-End Pipeline Blueprint for Research Paper Simplifier

1. Pre-processing Module:

The pre-processing module will include specialized steps tailored to the academic language:

- Text Normalization: Adapt normalization to academic content, ensuring formulas and citations are preserved.
- Tokenization: Develop a tokenizer that accurately handles complex academic nomenclature.
- Sentence Boundary Detection: Employ a model trained on academic texts for accurate sentence demarcation.
- Sentence Segmentation: Segment text into sentences for fine-grained analysis.
- Stopword Removal: Eliminate common stopwords to reduce noise.
- Lemmatization: Reduce words to their base forms to enhance model understanding.
- POS Tagging: Annotate words with their part-of-speech tags for context.
- Cleaning: Implement selective removal of non-textual elements, retaining important in-text references.
- Specialized Terminology Handling: Manage technical terms with dynamic lexicon updates and simplification strategies.

2. Definition of simplification (feature-wise):

Simplification will be approached through multiple dimensions:

- Sentence Structure Simplification: Transforming complex sentence structures into simpler forms.
- Lexical Simplification: Train models using the ASSET dataset for context-aware synonym replacement.
- Syntactic Simplification: Employ syntactic tree manipulations for structural simplification.
- Semantic Simplification: Use neural networks for abstracting complex arguments and rephrasing.

3. Scope of Simplification:

The simplification scope will address varying complexities:

- Complexity Levels: Adapt simplification to user profiles based on the ASSET dataset.
- Content Prioritization: Utilize algorithms for relevant content identification and simplification.

- Interactivity: Offer layered simplification and explanations upon user interaction.
- Simplifying technical jargon, reducing sentence length and complexity, making abstract concepts more understandable, preserving essential content while removing redundancies.

4. Problem Formulation:

Formulation will hinge on the ASSET dataset and unsupervised learning techniques:

- Inputs: Multi-paragraph texts with clear section delineation from research papers. (A research paper in text format.)
- Outputs: A simplified version of the research paper.

Loss Function:

- Output Activation Function: For the Transformer model engaged in sentence simplification, the softmax activation function will be employed at the output layer. The softmax function is particularly advantageous for multi-class classification problems, like word prediction, because it converts the model's raw output scores (logits) into probabilities that sum to one. This probabilistic output is essential for our task, as it provides a distribution over all possible next-word choices, facilitating the selection of the most probable word during sentence generation.
- Choice of Loss Function: Given the softmax output, categorical cross-entropy is an appropriate loss function. Categorical cross-entropy measures the distance between the probability distribution output by the softmax function and the actual distribution (the one-hot encoded vector representing the correct next word). In essence, it quantifies how well the probability distribution predicted by the model aligns with the distribution of the target data.
- Rationale for Using Categorical Cross-entropy: By minimizing this loss, the Transformer model is trained to increase the accuracy of the predicted probability distribution for the next word in the simplified sentence, which is essential for generating coherent and semantically appropriate text.

5. Methodology at Broad Level:

Unsupervised Neural Networks

The core of our methodology is based on the Transformer architecture, chosen for its state-of-the-art performance in various NLP tasks. The Transformer model excels in handling long-range dependencies and complex sentence structures, which are common in research papers.

- Transformer Architecture: The Transformer uses a self-attention mechanism that allows it to weigh the importance of different words in a sentence, regardless of their positional distance from each other. This is crucial for understanding which terms can be simplified while preserving the meaning of the sentence.

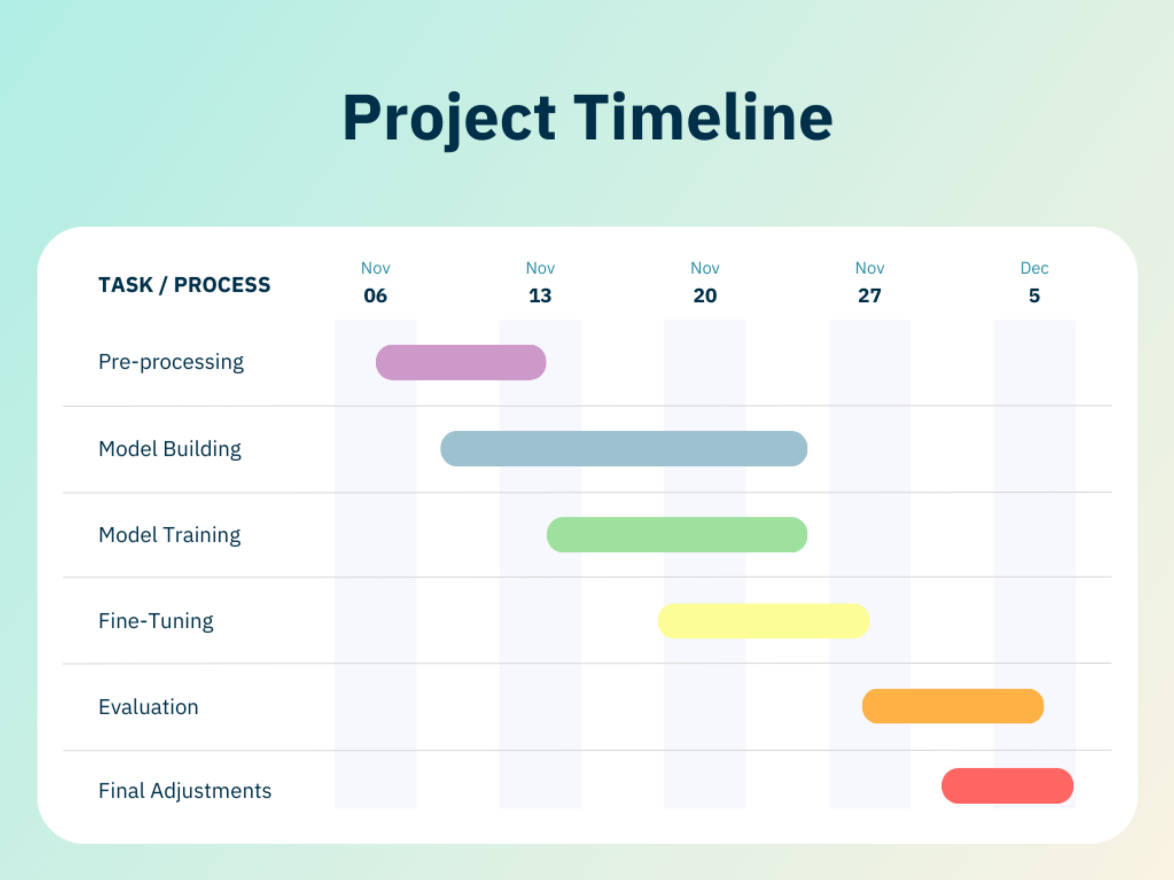
- Multi-Layer, Multi-Head Attention: We hypothesize that the multi-layer, multi-head attention mechanism of the Transformer is particularly well-suited for sentence simplification. It can simultaneously focus on different aspects of sentence complexity, such as syntactic structure and semantic nuances, and apply different simplification strategies within the same framework.

- Supervised Fine-Tuning: Although our primary approach is unsupervised, we plan to further enhance the Transformer model with supervised fine-tuning. This involves using the simplified sentences available in the ASSET dataset to refine the model's parameters, ensuring that the generated simplifications align more closely with human judgments of simplicity and clarity.

- Evaluation of Model Capability: We will conduct extensive evaluations to assess the model's capability in simplifying sentences. This will involve qualitative assessments by human judges as well as quantitative metrics such as BLEU, SARI, and FKGL to measure the readability and quality of the simplifications.

- Iterative Improvement: Based on the performance of the model, we will iteratively improve the architecture. This might include adjustments to the attention mechanisms, layer configurations, and training procedures to better capture the intricacies of sentence simplification.

6. Timeline for Completion of Project



7. Task Delegation Among Members:

Pre-processing / Evaluation : Bhargav Dave

Model Building/ Training/ Tuning : Purvi, Jigar

Team NeuronicNav

Purvi Patel (202211023)
Jigar Shekhat (202211004)
Bhargav Dave (202221004)